MRC Epidemiology Unit

Genetic Epidemiology in the New Era

November 4, 2008

Academy of Military Medical Sciences, Beijing, China

Jing Hua Zhao

MRC Epidemiology Unit, Cambridge, UK

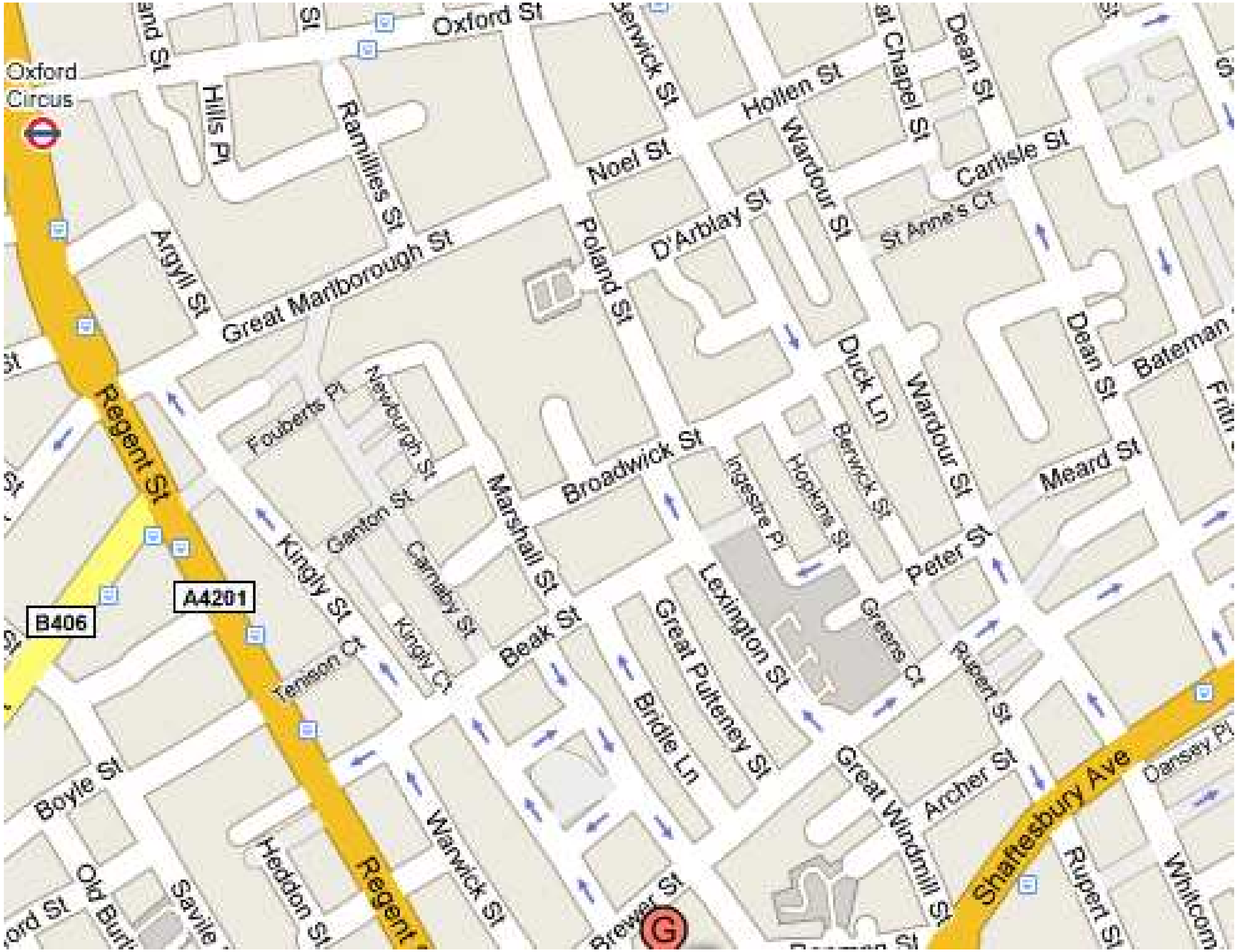
jinghua.zhao@mrc-epid.cam.ac.uk
http://www.mrc-epid.cam.ac.uk/~jinghua.zhao

Outline of presentation

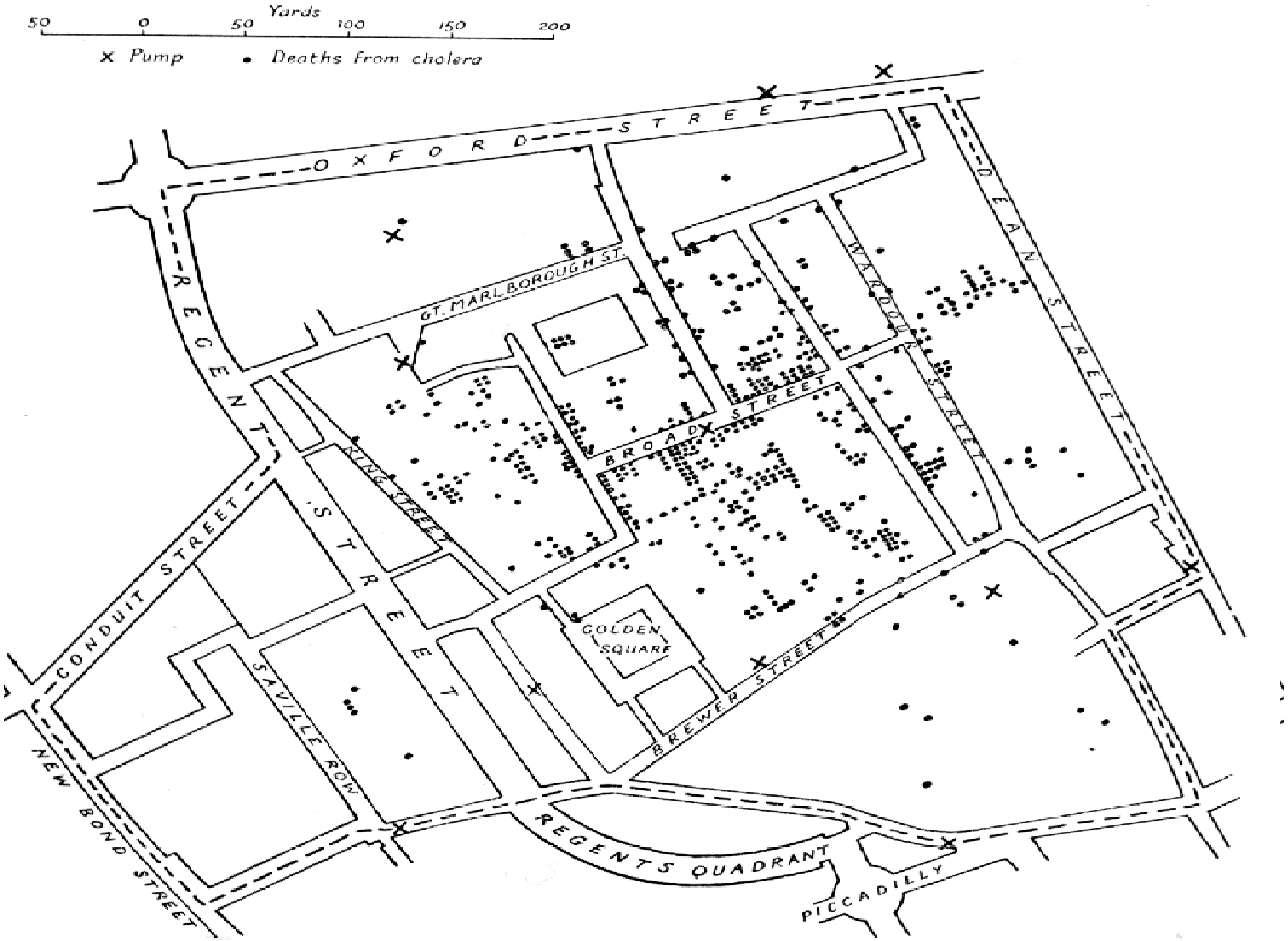
- I. Genetic epidemiology
    - I.1 Analysis of familial aggregation
    - I.2 Segregation analysis
    - I.3 Linkage analysis
    - I.4 Association analysis
- II. Genome epidemiology
    - II.1 GAW15 expression data analysis
    - II.2 GWAS of anthropometric and related traits
- III. Gene characterization
    - III.1 *APOE*, socioeconomic status and cognitive function
    - III.2 GAW16 FHS data analysis
- IV. Summary

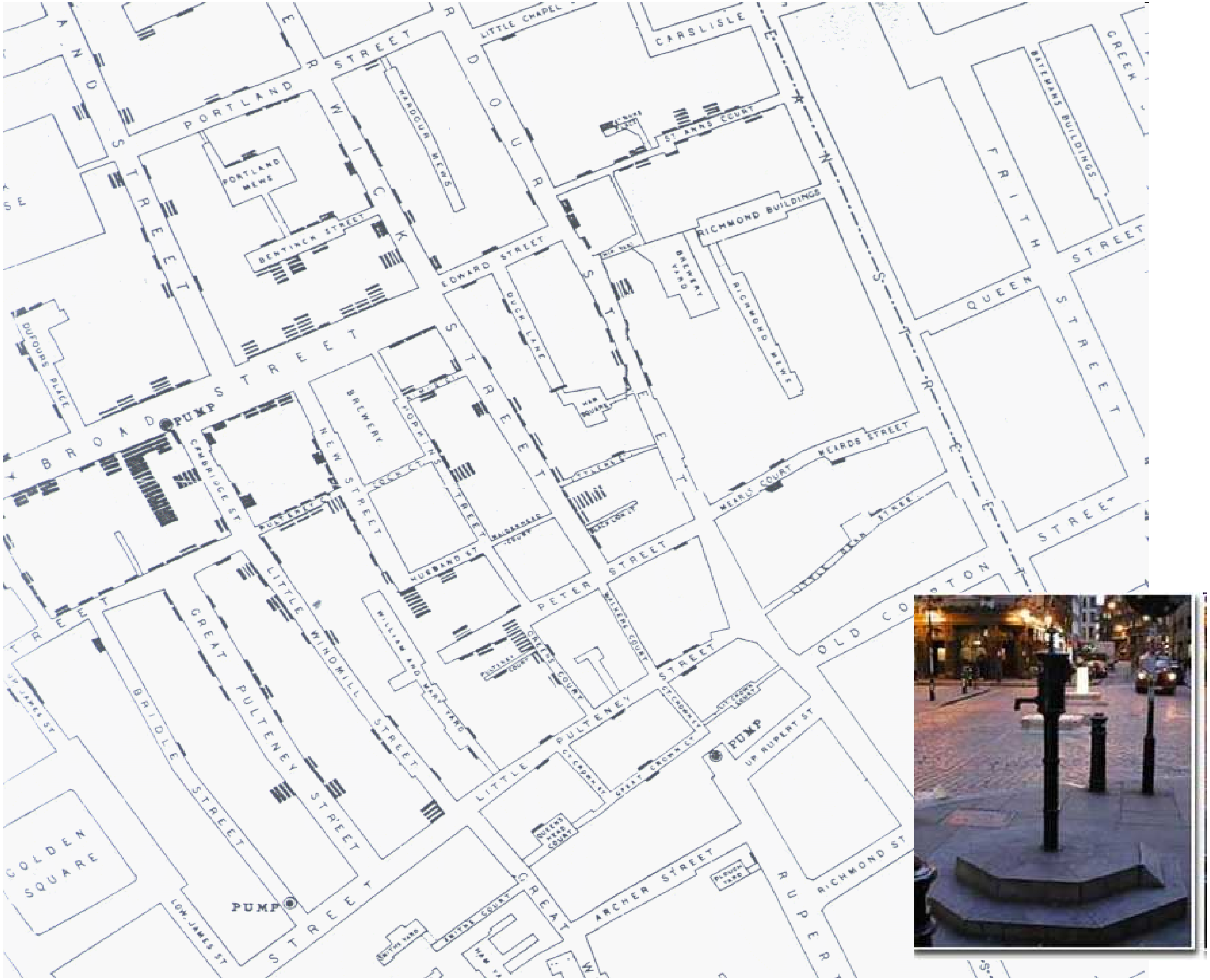
I. Genetic Epidemiology

Broad street as of today (Google maps)



Epidemiology: the 1854 London cholera epidemic





http://www.csiss.org/classics/

Definition (Wikepedia)

The study of the role of genetic factors in determining health and disease in families and in populations, and the interplay of such genetic factors with environmental factors. Slightly more formally, genetic epidemiology was defined by Morton as "*a science which deals with the etiology, distribution, and control of disease in groups of relatives and with inherited causes of disease in populations*“ (Morton. Outline of Genetic Epidemiology, 1982).

It is closely allied to molecular epidemiology, a branch of public health that deals with the contribution of potential genetic and environmental risk factors identified at the molecular level, to the etiology, distribution and control of the disease in groups of relatives and populations. It improves our understanding of the pathogenesis of disease by identifying specific pathways, molecules and genes that influence the risk of developing disease.

The emphasis here, however, is on the statistical aspects.

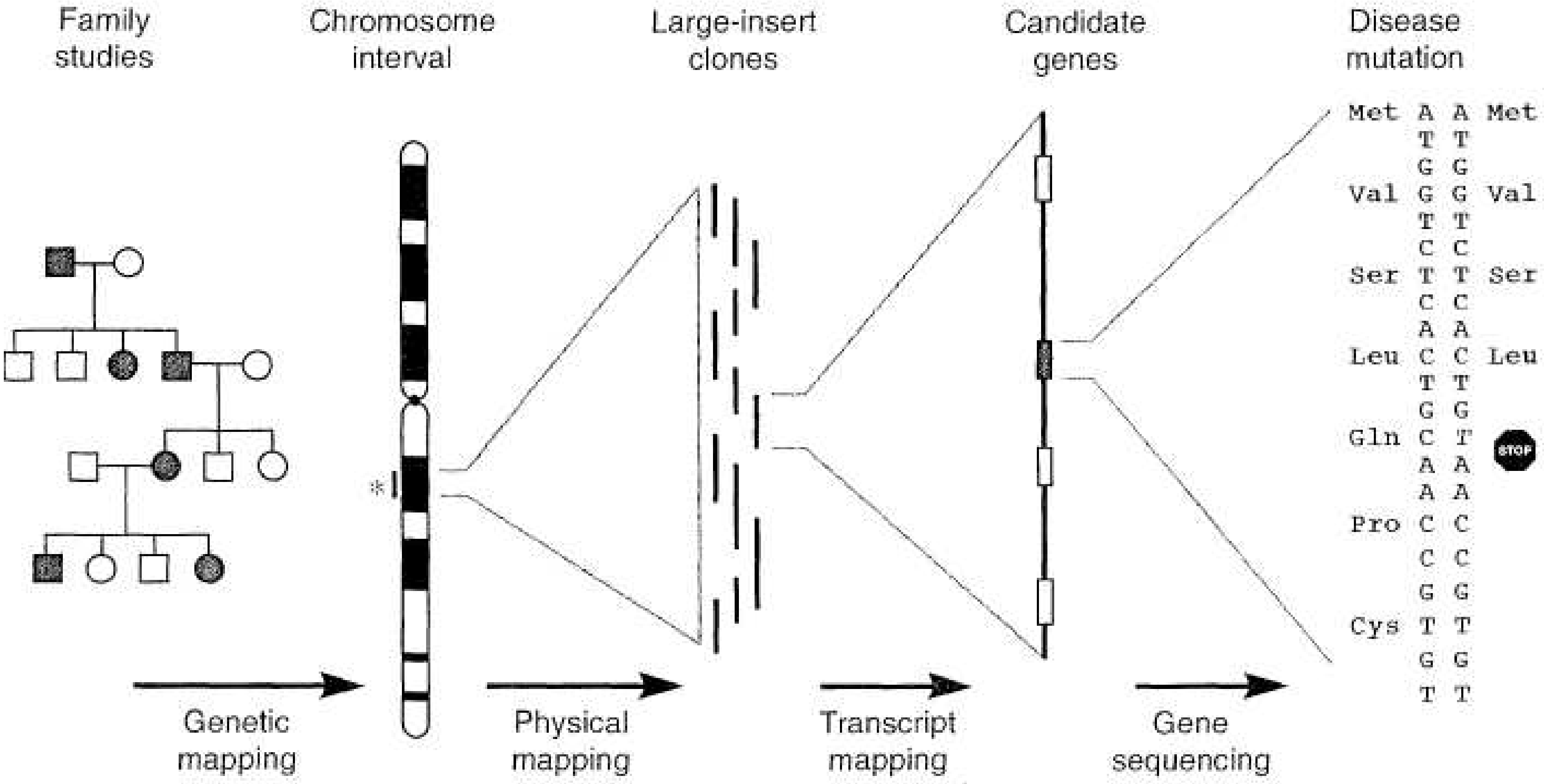
Complex traits (Lander & Schork 1994)

- **Incomplete penetrance**: the probability that individuals inheriting the gene will have the disease is less than 1 and dependent on factors such as age
- **Phenocopy**: a disease due to nongenetic causes
- **Genetic heterogeneity**: a disease due to different genetic mutations to different individuals
- **Polygenic inheritance**: the liability of disease due to additive or interactive effects at multiple loci
- Other phenomena such as **imprinting**, the difference in function of an allele according to its parental origin

Paradigms (Wikepedia)

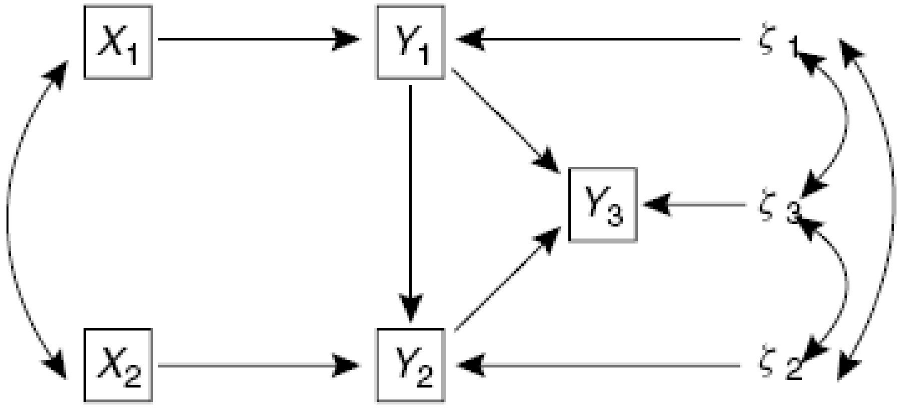
- Familial aggregation: Is there a genetic component to the disease, and what are the relative contributions of genes and environment?
- Segregation: What is the pattern of inheritance of the disease (e.g. dominant or recessive)?
- Linkage: On which part of which chromosome is the disease gene located?
- Association: Which allele of which gene is associated with the disease?

Steps for positional cloning (Schuler et al. 1996)



I.1 Analysis of Familial Resemblance

Path analysis (Bollen 2005)



$$Y_1 = \alpha_1 + \gamma_{11} X_1 + \zeta_1,$$

$$Y_2 = \alpha_2 + \beta_{21} Y_1 + \gamma_{22} X_2 + \zeta_2,$$

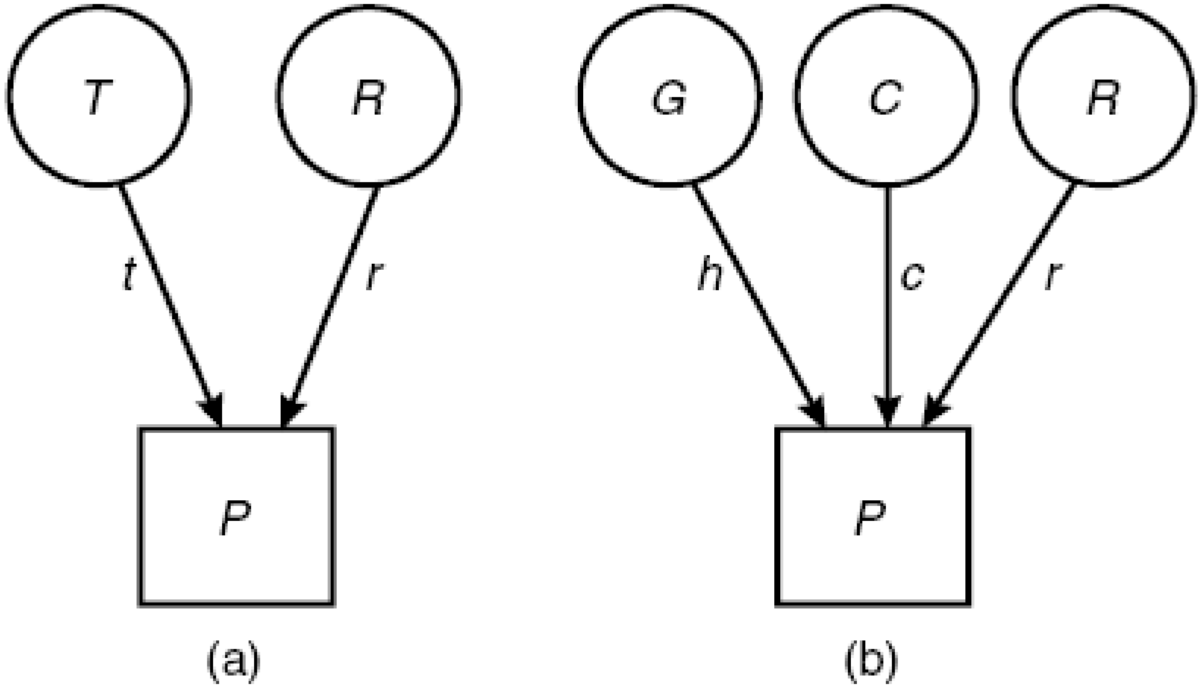
$$Y_3 = \alpha_3 + \beta_{31} Y_1 + \beta_{32} Y_2 + \zeta_3,$$

$$\mathrm{cov}(x_k, \zeta_i), \qquad \mathrm{cov}(\zeta_i, \zeta_j) \neq 0,$$

$$\mathrm{E}(\zeta_i) = 0, \qquad i, j = 1, 2, 3 \text{ for } i \neq j, k = 1, 2.$$

- A simultaneous equations model, with the relationship between variables specified as a path diagram and in mathematical form

Basic models (Rao & Rice 2005)

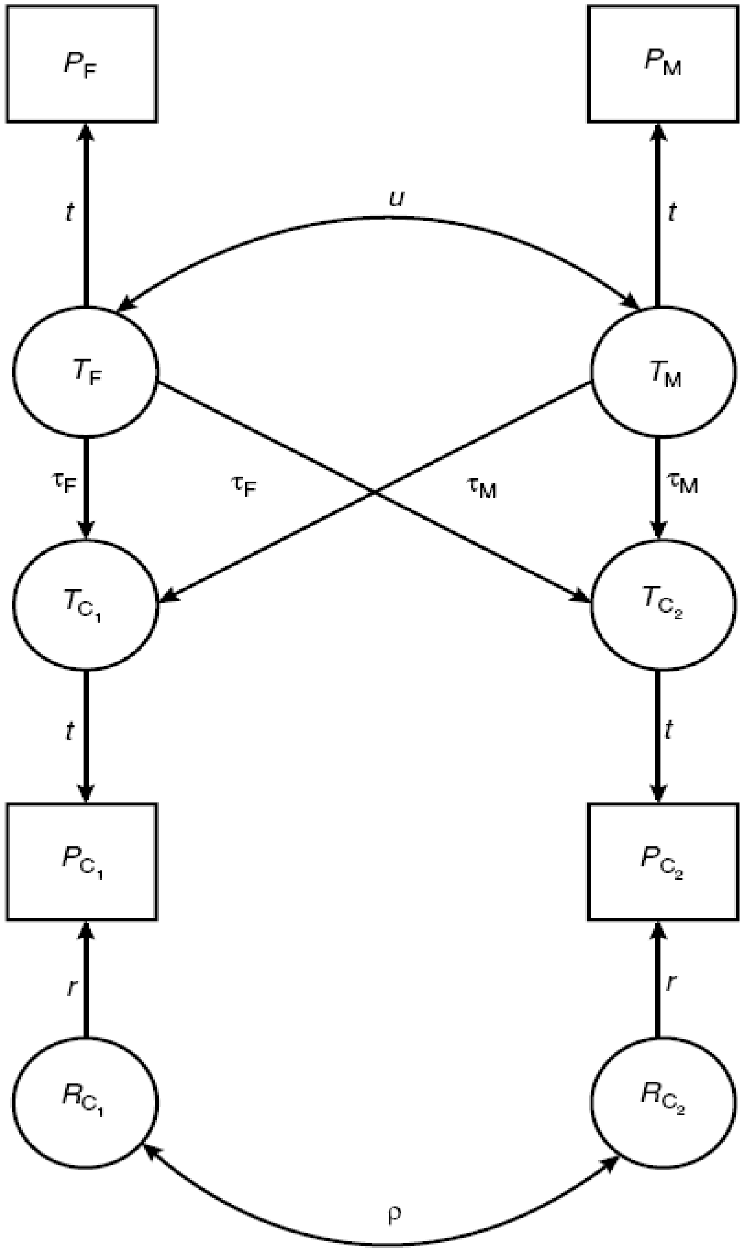


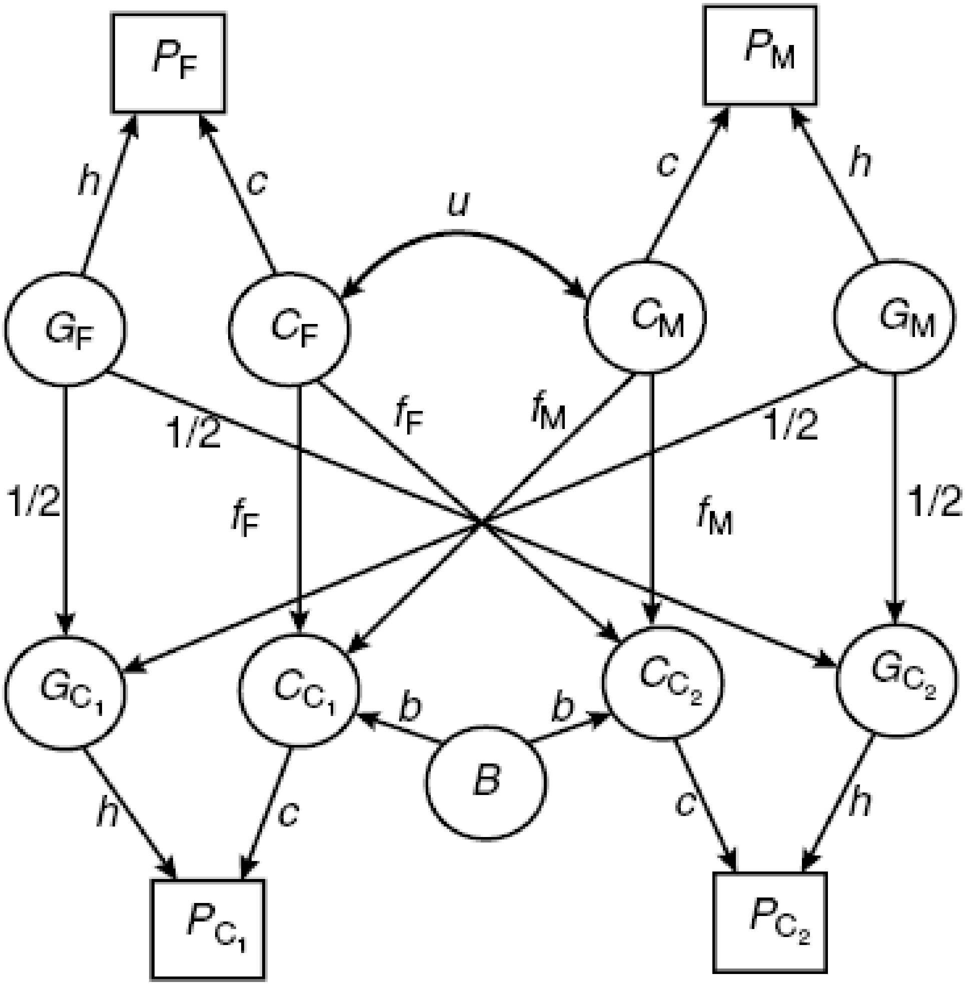
- Phenotypic variation (P) in an individuals is specified by two models: (a) a combined familial component (T) and a residual (T); (b) latent genetic (G), familial environmental (C), and residual (R).

Basic models

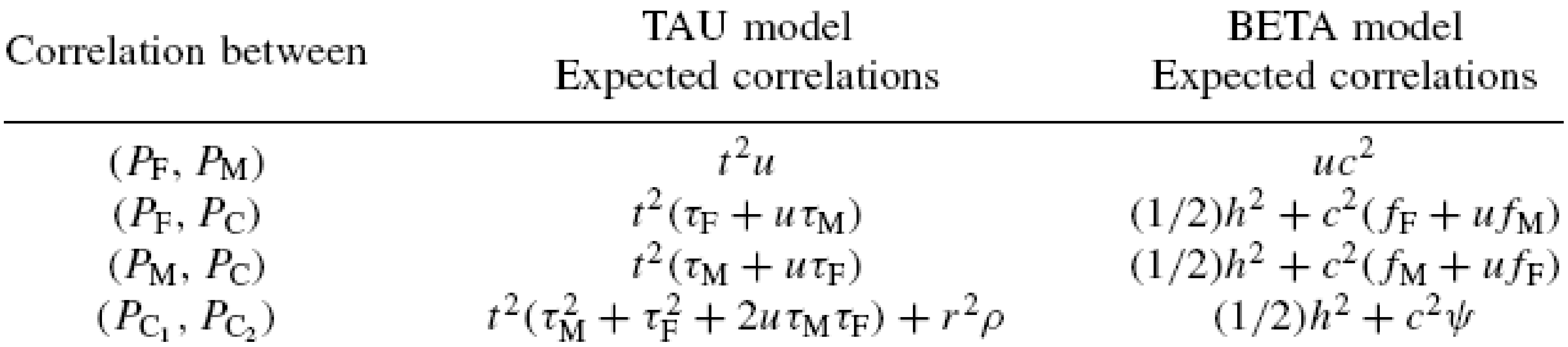
- Denote the unstandardized mean-centered variables with an asterisk, we have (a) P*=T*+R* and (b) P*=G*+C*+R* and the underlying structural equations (a) P=tT+rR and (b) P=hG+cC+rR, where the capital and lower letters represent standardized variables and path coefficients, respectively.
- This provides partition of the total phenotypic variance, of interest would be proportion of familial or environmental components relative to the total phenotypic variance, i.e., the heritability h^2 and the cultural heritability c^2.
- The appropriate path diagrams for nuclear family data are called TAU and BETA models.

TAU and BETA models





Model parameterization and inference



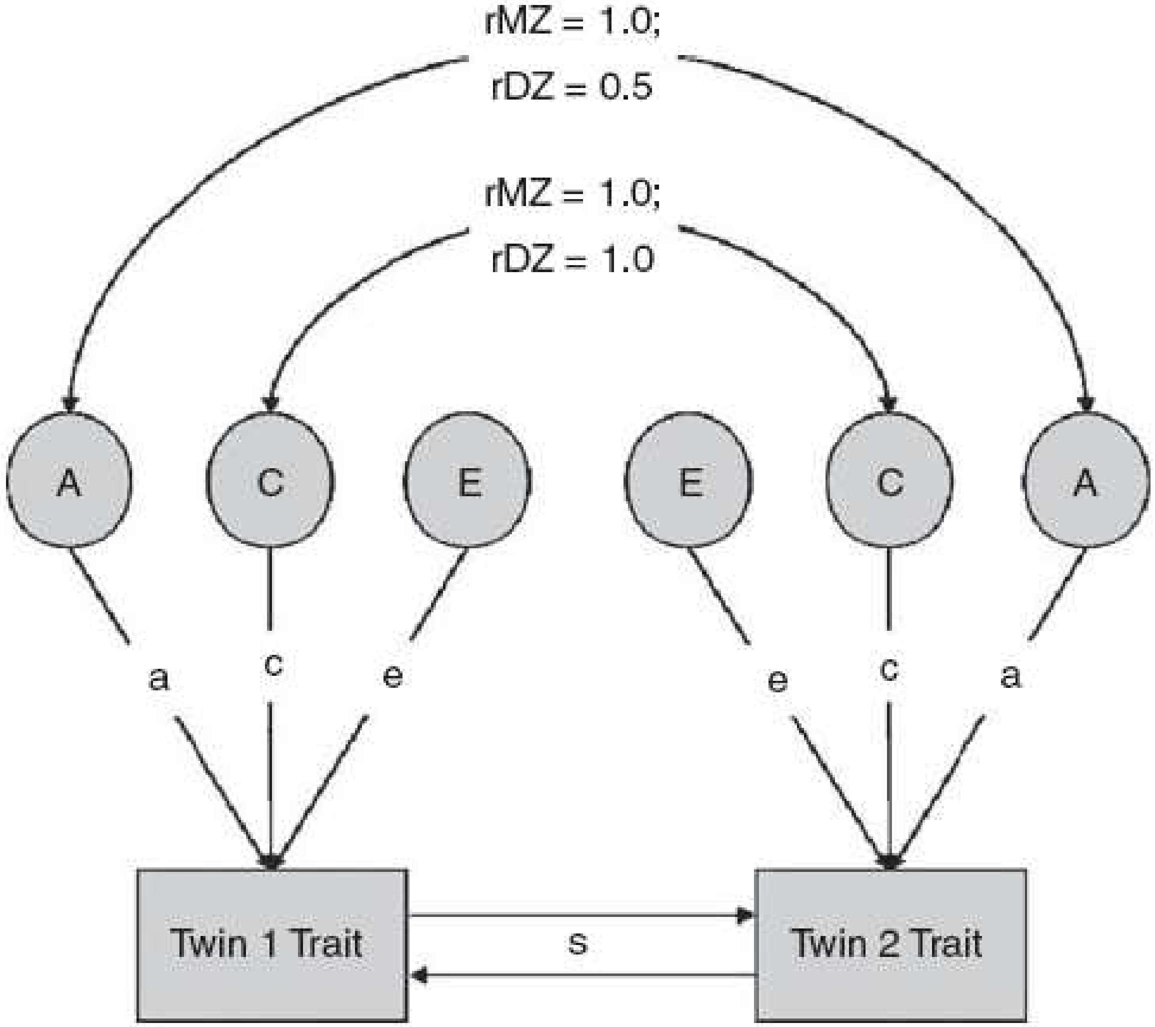
Correlation between	TAU model Expected correlations	BETA model Expected correlations
(P_F, P_M)	$t^2 u$	uc^2
(P_F, P_C)	$t^2(\tau_F + u\tau_M)$	$(1/2)h^2 + c^2(f_F + uf_M)$
(P_M, P_C)	$t^2(\tau_M + u\tau_F)$	$(1/2)h^2 + c^2(f_M + uf_F)$
(P_{C_1}, P_{C_2})	$t^2(\tau_M^2 + \tau_F^2 + 2u\tau_M\tau_F) + r^2\rho$	$(1/2)h^2 + c^2\psi$

$$\psi = b^2 + f_F^2 + f_M^2 + 2uf_F f_M$$

The correlation between relatives can be parameterized into a MVN(μ,Σ) so that statistical inference can be made based on the log-likelihood function.

$$\ln L = -\left(\frac{1}{2}\right)[\ln|\boldsymbol{\Sigma}| + (\mathbf{X} - \mu)'\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mu)]$$

Appetite in children (Carnell et al. 2008)

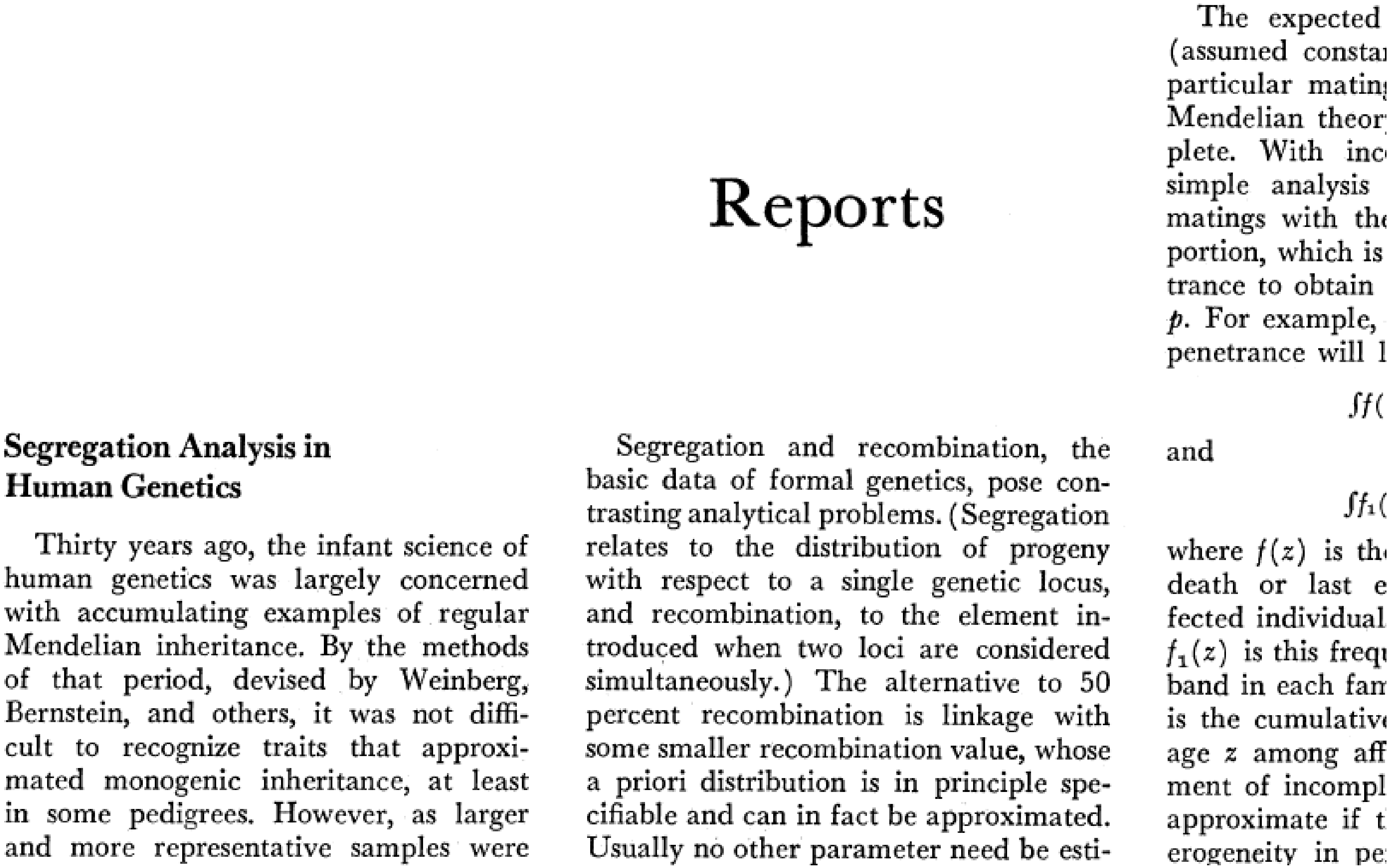


Study sample and heritability estimates

- A population-based sample of 8- to 11-year-old 5435 twin pairs in the Twins' Early Development Study (TEDS)
- Heritability estimates are 0.63 (95%CI, 0.39-0.81) for satiety responsiveness, and 0.75 (95%CI 0.52-0.85) for food cue responsiveness, shared and nonshared environmental influences were 0.21(0-0.51) and 0.16 (0.1-0.21) for satiety responsiveness and 0.1 (0-0.38) and 0.15 (0.1-0.18), respectively.
- It was concluded that genetic vulnerability to weight gain could operate through behavioral as well as metabolic pathways.

II.2 Segregation Analysis

Segregation analysis (Morton 1958)



Reports

Segregation Analysis in Human Genetics

Thirty years ago, the infant science of human genetics was largely concerned with accumulating examples of regular Mendelian inheritance. By the methods of that period, devised by Weinberg, Bernstein, and others, it was not difficult to recognize traits that approximated monogenic inheritance, at least in some pedigrees. However, as larger and more representative samples were

Segregation and recombination, the basic data of formal genetics, pose contrasting analytical problems. (Segregation relates to the distribution of progeny with respect to a single genetic locus, and recombination, to the element introduced when two loci are considered simultaneously.) The alternative to 50 percent recombination is linkage with some smaller recombination value, whose a priori distribution is in principle specifiable and can in fact be approximated. Usually no other parameter need be esti-

The expected
(assumed constai
particular matin;
Mendelian theor;
plete. With inc
simple analysis
matings with th
portion, which is
trance to obtain
p. For example,
penetrance will l

$\int f($

and

$\int f_1($

where $f(z)$ is th
death or last e
fected individual
$f_1(z)$ is this frequ
band in each fam
is the cumulativ
age z among aff
ment of incompl
approximate if t
erogeneity in pe

Mixed models

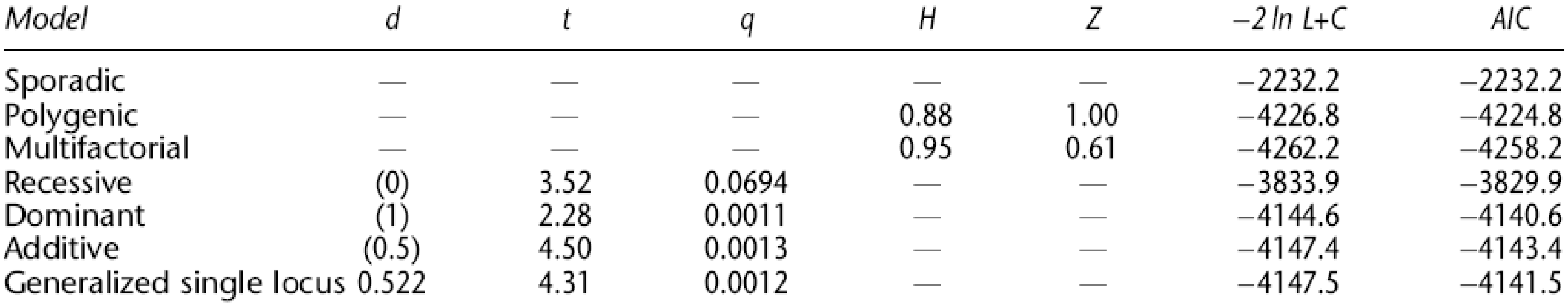
- A phenotype of interest (x) is a result of major locus, polygenic and environmental effects
- The genotypic effects of AA, Aa and aa can be characterized by
  - z = $-(tq^2 + 2pqdt)$
  - t: $t^2 = 1/[(pq)^2(q + 2pd)^2 + 2pq(d - q^2 - 2pqd)^2 + q^2(1 - q^2 - 2pqd)^2]$
  - q, d = frequency of a, and displacement; they are the free parameters

z	*z+dt*	*z+t*
AA	Aa	aa

- POINTER obtains maximum likelihood estimators by iterating mean and variance of x, d, q, H (polygenic heritability) and B (relative variance due to common environment)

Nasopharyngeal carcinoma (Jia et a. 2005)

- A total of 1903 Cantonese pedigrees partitioned into 3737 nuclear families. A mixed as implemented in POINTER shows no evidence for a major gene and the observed data are best explained by a multifactorial mode of inheritance.
- d: where d=0 refers to a recessive major gene, d=0.5 additive and d=1 dominant; t: difference on the liability scale between homozygotes; q: gene frequency; H: the polygenic heritability; Z: the ratio of the multifactorial component in adults and children; and AIC: Akaike’s Information Criterion.

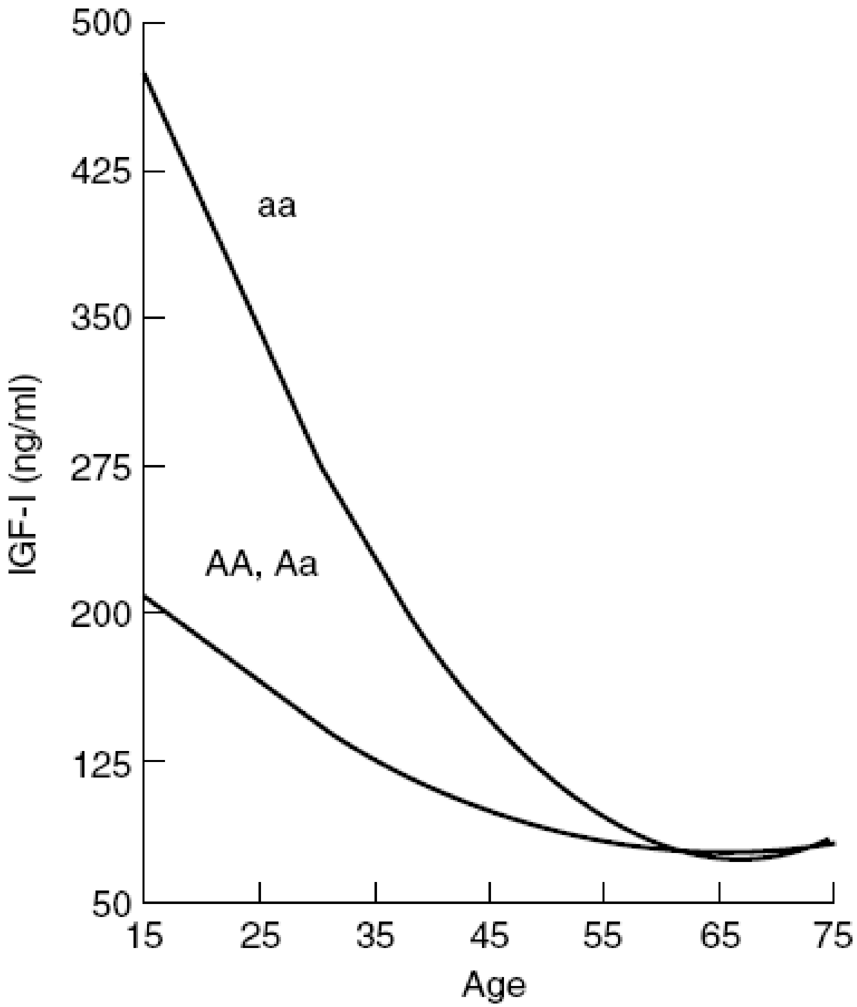


Model	*d*	*t*	*q*	*H*	*Z*	*−2 ln L+C*	*AIC*
Sporadic	—	—	—	—	—	−2232.2	−2232.2
Polygenic	—	—	—	0.88	1.00	−4226.8	−4224.8
Multifactorial	—	—	—	0.95	0.61	−4262.2	−4258.2
Recessive	(0)	3.52	0.0694	—	—	−3833.9	−3829.9
Dominant	(1)	2.28	0.0011	—	—	−4144.6	−4140.6
Additive	(0.5)	4.50	0.0013	—	—	−4147.4	−4143.4
Generalized single locus	0.522	4.31	0.0012	—	—	−4147.5	−4141.5

Analysis of IGF1 in Mexican Americans

A study of 422 Mexican Americans from 24 pedigrees for insulin-like growth factor 1 (IGF1) showed evidence of a major gene influencing IGF1 levels. The estimated frequency of the A allele associated with lowered IGF1 levels was 0.54+/-0.05. Log-likelihood ratio tests revealed evidence for genotype by age interaction. Individuals with genotypes AA and Aa showed a less marked decline in IGF1 levels with age than that observed for the aa genotype.

(Blangero 2005)



Regressive models

- Distributions over pedigrees are specified by conditioning each individual’s trait value on those of antecedent individuals.
- For binary traits, a multivariate logistic model is formed
- For continuous trait, they assume (after transformation when appropriate) multivariate normality across members of the individual residuals from the type means.
  - Class A models assume sibling subtypes are dependent only because of common parentage
  - Class D models assume that the sibling correlations are equal but not necessarily due to common parentage alone.
- For both continuous and binary traits, finite polygenic mixed model can be used, including binary traits with variable age of onset.

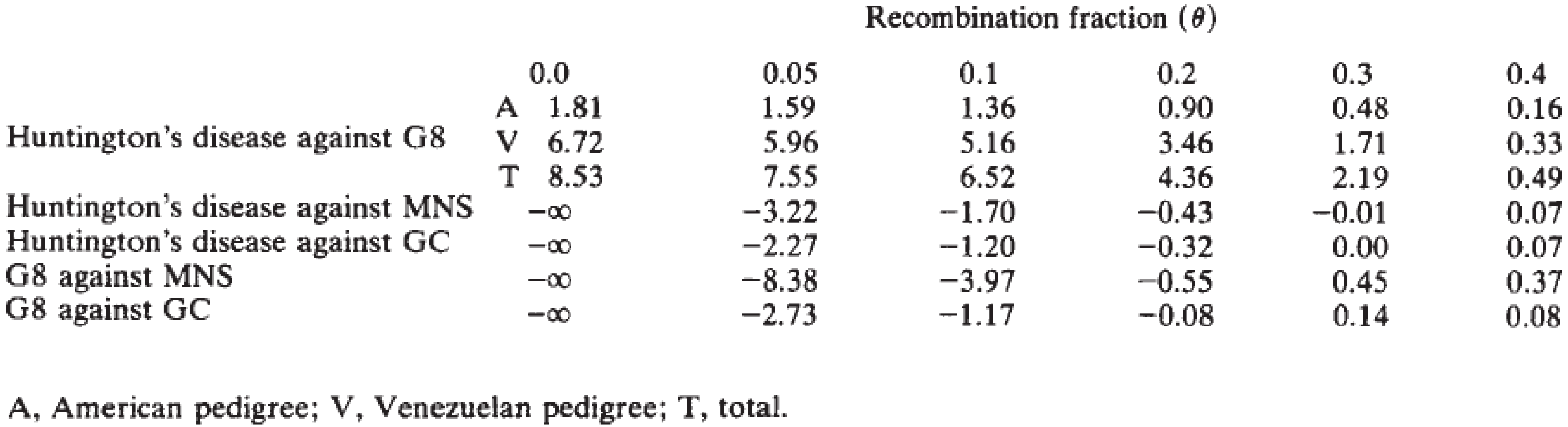
Breast cancer

- Claus et al. (1991) showed a dominant major gene model is more preferable to pure environmental, pure polygenic, or recessive major gene models
- Andrieu & Damenais (1997) using Bonney’s class D model showed women with a young age at menarche have dramatically higher penetrances than those with older ages at menarche
- This led to the finding of *BRCA1* and *BRCA2* genes
- A breast and ovarian analysis of disease incidence and carrier estimation algorithm (BOADICEA) is available http://www.srl.cam.ac.uk/genepi/boadicea/boadicea_home.html as with R/Bayesmendel, http://www.cancerbiostats.onc.jhmi.edu/BayesMendel

(Thomas (2004) *Statistical Methods in Genetic Epidemiology* and Blangero J (2005) in Encyclopedia of Biostatistics II)

I.3 Linkage Analysis

Linkage of Huntington’s disease



		Recombination fraction (θ)					
		0.0	0.05	0.1	0.2	0.3	0.4
	A	1.81	1.59	1.36	0.90	0.48	0.16
Huntington’s disease against G8	V	6.72	5.96	5.16	3.46	1.71	0.33
	T	8.53	7.55	6.52	4.36	2.19	0.49
Huntington’s disease against MNS		$-\infty$	−3.22	−1.70	−0.43	−0.01	0.07
Huntington’s disease against GC		$-\infty$	−2.27	−1.20	−0.32	0.00	0.07
G8 against MNS		$-\infty$	−8.38	−3.97	−0.55	0.45	0.37
G8 against GC		$-\infty$	−2.73	−1.17	−0.08	0.14	0.08

A, American pedigree; V, Venezuelan pedigree; T, total.

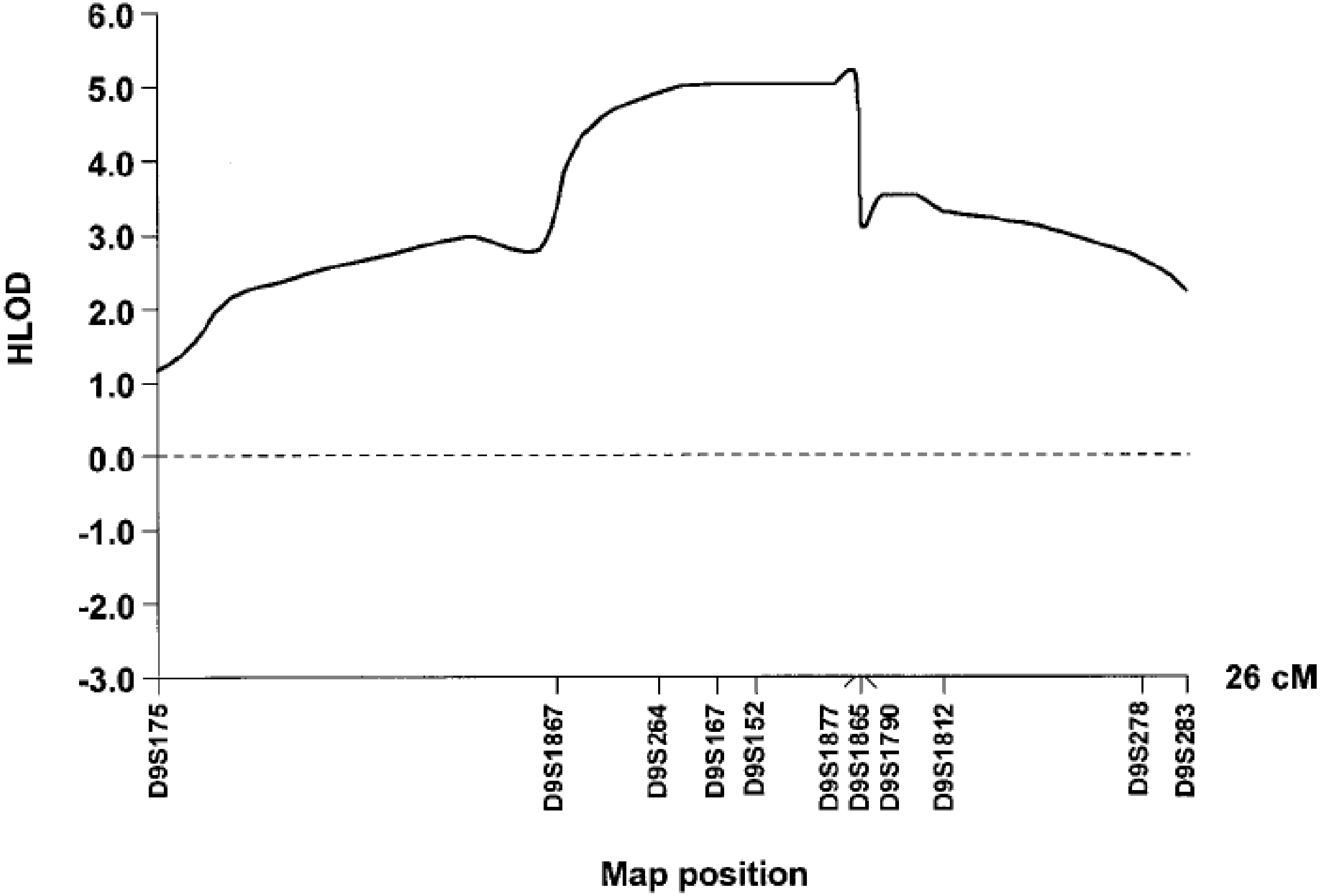
- The lod score between the Huntington’s disease locus and G8 locus at chromosome 4 is 6.52, with 99%CI 0-10cM. However there is no evidence of linkage with MNS and GC loci. (Gusella et al. 1983)

Familial hemophagocytic lymphohistiocytosis

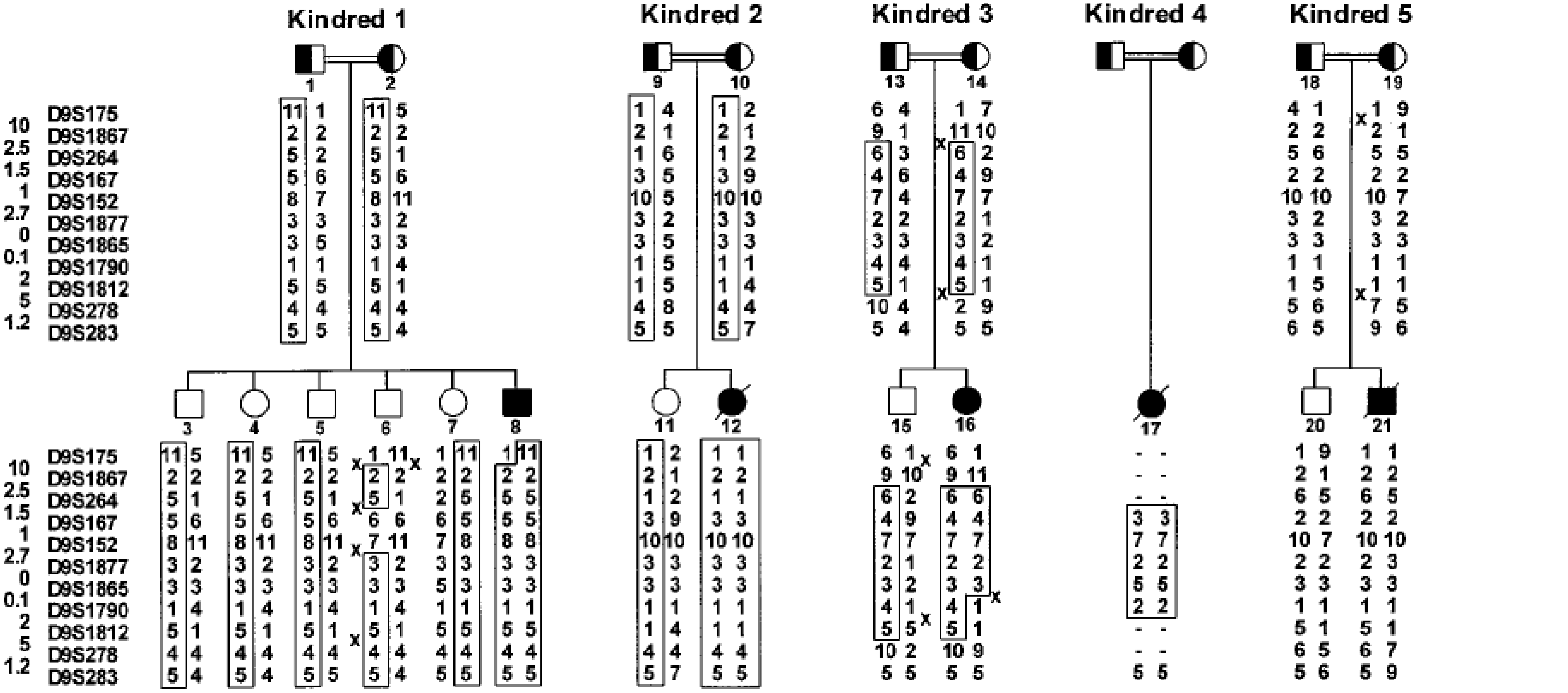
(Ohadi et al. 1999)

- An example of homozygosity mapping involving an autosomal recessive model with complete penetrance and a frequency of 0.004 for the disease allele
- Power was estimated by SLINK. Lod scores of 4.61 and 3.67 were obtained for markers 5 and 10 cM from the disease gene assuming five equi-frequent alleles.
- Genotyped done at the HGMP-RC.
- Allele frequencies were estimated from 50 unrelated, randomly selected, healthy Pakistani subjects. Analysis was done with GENEHUNTER. A MLOD of 4.40 and HLOD of 5.00 with α=0.81, at interval D9S1867-D9S1790. The admixture chi-squared test of heterogeneity gave a value of 2.76 and significant level 0.1.

HLOD of the five FHL families



Allele segregation of linked and flanking markers



I.4 Association Analysis

II. Genome Epidemiology

Derivation from HuGENet

- … the assessment of the impact of human genome variation on population health & how genetic information can be used to improve health & prevent disease

GAW (http://www.gaworkshop.org)

The Genetic Analysis Workshops (GAWs) are a collaborative effort among genetic epidemiologists to evaluate and compare statistical genetic methods. For each GAW, topics are chosen that are relevant to current analytical problems in genetic epidemiology, and sets of real or computer-simulated data are distributed to investigators worldwide. Results of analyses are discussed and compared at meetings held in even-numbered years.

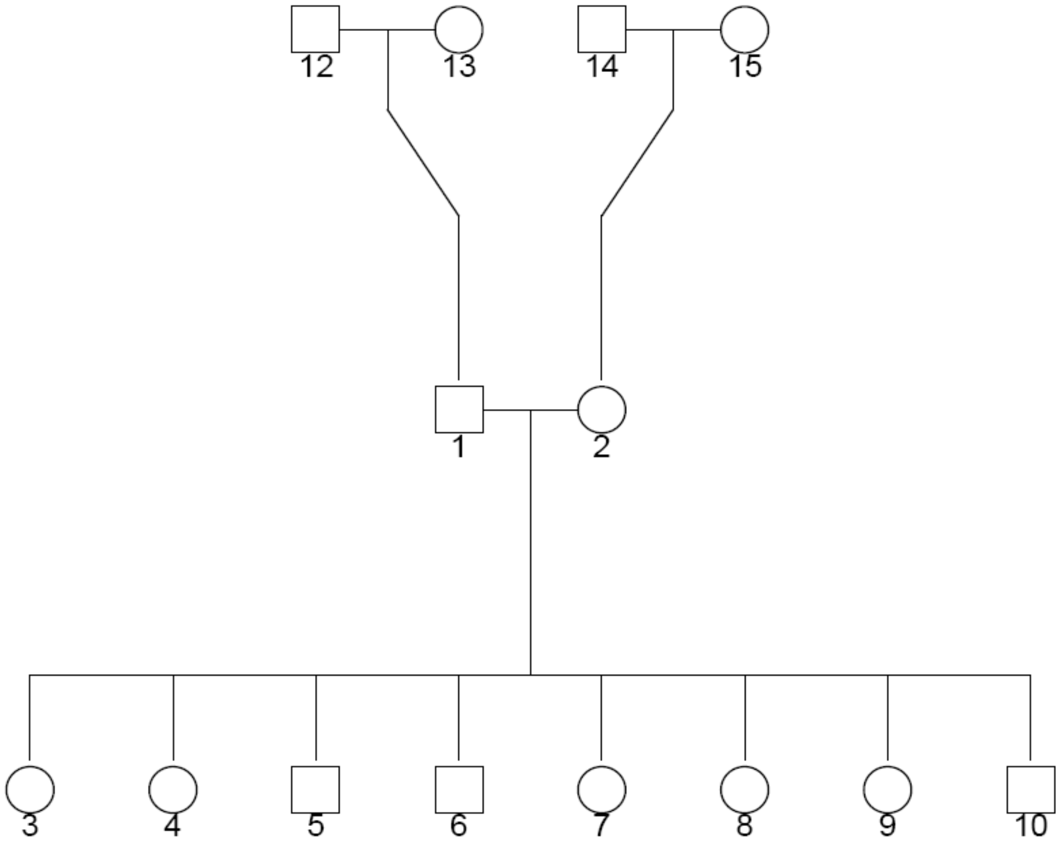
II.1 GAW15
Gene expression analysis

Background

The Genetic Analysis Workshop 15 (GAW15) Problem 1 contained baseline expression levels of 8793 genes in immortalized B cells from 194 individuals in 14 Centre d'Etude du Polymorphisme Humain (CEPH) Utah pedigrees. Previous analysis of the data (Morley et al**.** Nature 2004, **430:** 743-74) showed linkage and association and evidence of substantial individual variations. In particular, correlation was examined on expression levels of 31 genes and 25 target genes corresponding to two master regulatory regions.

Gene expression levels and aging (Tan et al. 2008)

Microarray gene expression data in 194 individuals from 14 CEPH Utah families were obtained by hybridising RNA extracted from immortalised lymphoblastoid cells to the Affymetrix focus array containing ~8500 genes, the analysis focused on demographic characteristics such as age and sex on differential gene expression patterns.



Analysis

- Normalisation was performed for pre-processing data.
- The major statistical method is GEE, which accounts for within individual correlation. In this technique one has the flexibility to specify correlation structures. It has been implemented in R (as well as Stata, SAS). To simplify the analysis, the parent generation is dropped from the analysis. The sibling correlation is also quantified with intraclass correlation (ICC), empirical significance is assessed by resampling method implemented in R.
- To account for multiple-testing, the false-discovery rate (FDR) is used.
- Hierarchical cluster analysis is used for the significant genes with *gplot* package.
- EASE software (http://david.abcc.ncifcrf.gov/ease/ease.jsp) is used to cluster significantly regulated genes into pathways.
- The Affymetrix Focus database was used to extract information on gene annotation.

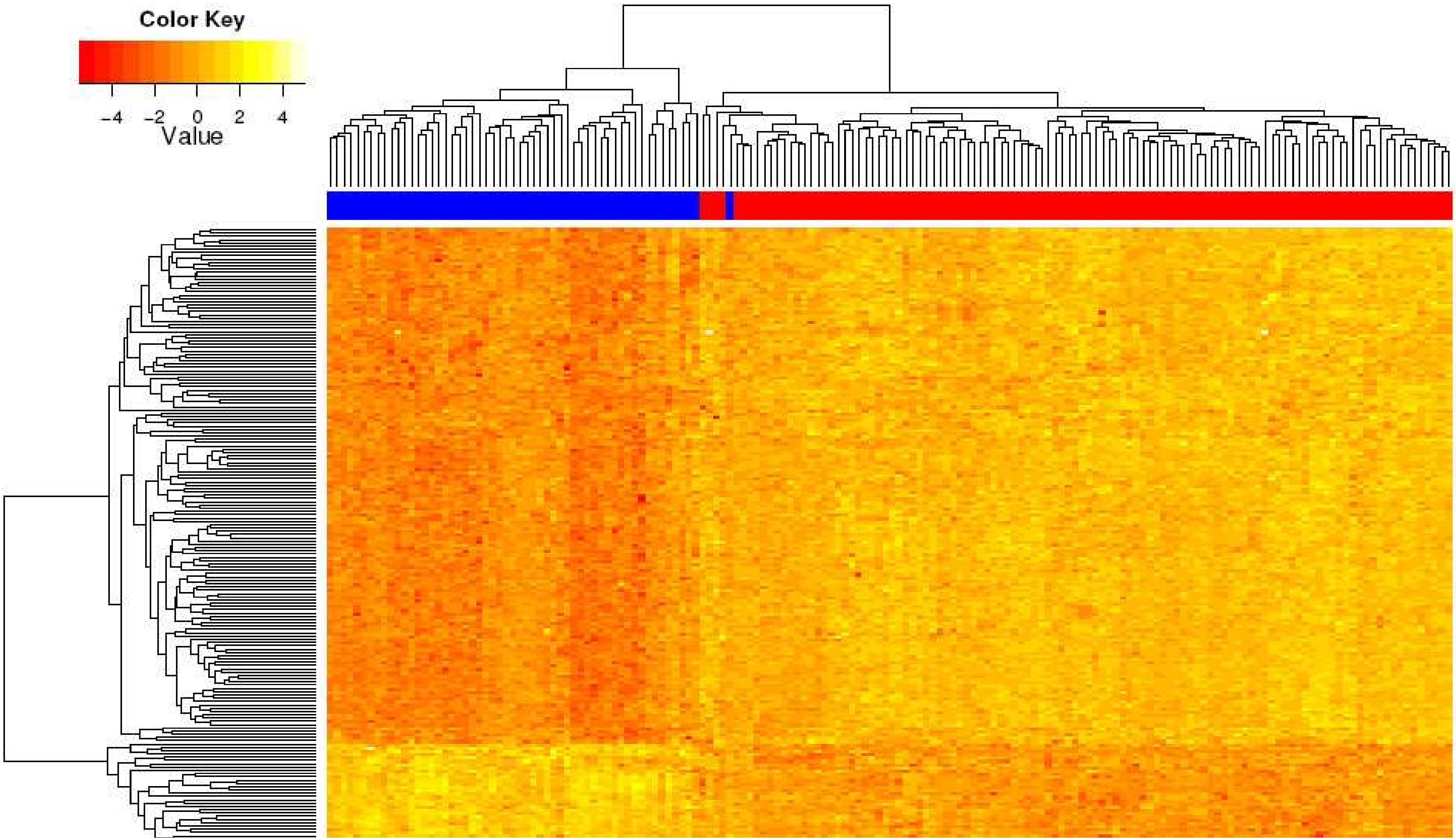
200 topmost significant genes

Gene Category	List hits	Pop. hits	score
Cell-cell signaling	30	495	0.000
Cell communication	68	2402	0.008
Inorganic anion transport	6	60	0.009
Channel/pore class transporter activity	14	290	0.010
Chloride transport	5	42	0.012
Signal transducer activity	51	1733	0.014
Voltage-gated ion channel activity	8	119	0.014
Alpha-type channel activity	13	280	0.018
Anion transport	7	98	0.019
Extracellular	32	979	0.022
Ion channel activity	12	258	0.024
Ion transport	17	433	0.025
Muscle contraction	8	136	0.028
Development	42	1422	0.028
Receptor binding	17	441	0.028
Cell surface receptor linked signal transduction	28	875	0.036
Sex differentiation	3	14	0.036
Organogenesis	26	804	0.039
Voltage-gated chloride channel activity	3	15	0.040
Transcription factor complex	17	458	0.043
Enzyme linked receptor protein signaling pathway	8	151	0.045
TGFbeta receptor signaling pathway	4	38	0.048

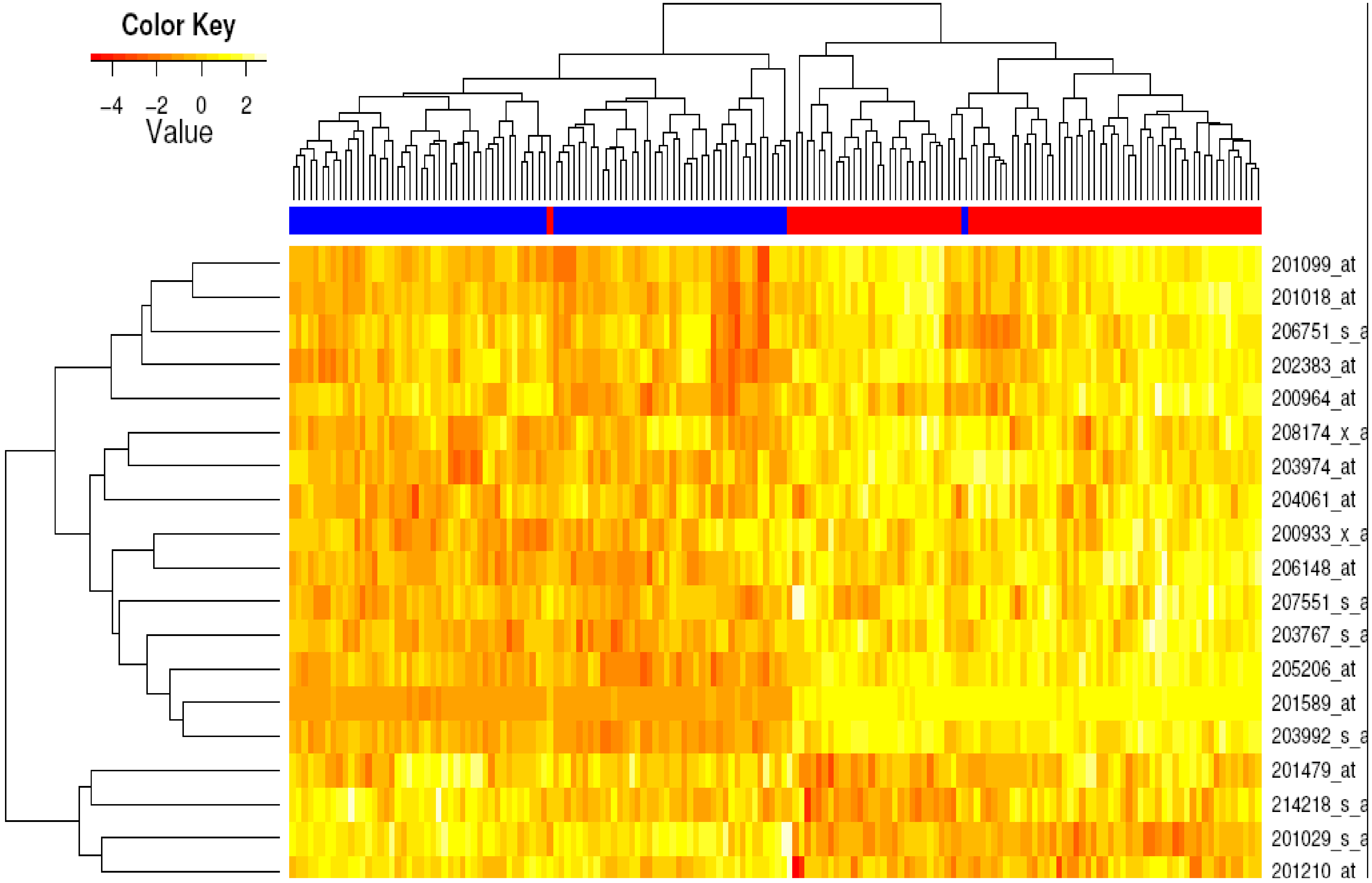
101 significant genes regulated by sex in the young

System	Gene Category	List hits	Population hits	EASE score
Molecular Function	RNA binding	15	375	0.000
Molecular Function	Nucleic acid binding	35	1897	0.002
Cellular Component	Ribonucleoprotein complex	10	290	0.006
Biological Process	RNA processing	9	248	0.007
Cellular Component	Spliceosome complex	5	69	0.008
Biological Process	RNA metabolism	9	269	0.012
Cellular Component	cAMP-dependent protein kinase complex	3	15	0.013
Cellular Component	Obsolete cellular component	10	324	0.013
Biological Process	RNA splicing	5	92	0.021
Biological Process	Development	25	1422	0.027
Biological Process	Male gamete generation	5	102	0.029
Biological Process	Spermatogenesis	5	102	0.029
Biological Process	Sexual reproduction	6	152	0.030
Molecular Function	Pre-mRNA splicing factor activity	4	61	0.030
Biological Process	Reproduction	6	153	0.030
Cellular Component	Nucleus	34	2101	0.032
Molecular Function	Cell adhesion molecule activity	8	277	0.035
Molecular Function	Chromatin binding	3	29	0.041

200 topmost significant genes regulated by age (Nearly all subjects are clearly distinguished left=grandparents, right=grandchildren)



The top 19 X-linked genes (Nearly all young males are to the left and females to the right panel)



Pathway analysis (Zhao et al. 2007)

- If the expression level of a given gene is regulated by certain proteins then it should be a function of the active levels of these proteins. Due to biological variability and measurement errors, the function would be stochastic rather than deterministic.
- Expression levels of genes are proxies for the activity level of the proteins they encode, although there are numerous examples where activation or silencing of a regulator is carried out by post-transcriptional protein modifications.

Statistical model

- Gene expression levels, treated as continuous variables, can be assumed to follow a multivariate normal distribution, and to be consistent with a Bayesian network with linear Gaussian conditional densities.
- The prior of this network is characterised by a prior network reflecting our belief in the joint distribution of the variables in question, and equivalent sample size (ESS) effectively behaving as if it was calculated from a “prior” data set of that size. For instance, without a priori knowledge of the regulatory network, the prior network could be one where all expression levels are independent in order to avoid explicitly biasing the learning procedure to a particular edge.
- The common approach to the learning procedure starts with a training set and evaluates networks according to an asymptotically consistent scoring function that is obtained through the Bayesian framework.

Statistical analysis

- Affymetrix CEL-files were preprocessed with BioConductor package *affy,* but the target gene expressions were used directly. The probe set IDs were matched with the annotation database of human genome focus array distributed with GAW15 problem 1 and from the Affymetrix (http://www.affymetrix.com). All data management, correlation and hierarchical cluster analysis were done with the R system (http://www.r-project.org)
- In the case of B-course software (http://b-course.hiit.fi), discretization of continuous data has been applied to capture the nonlinear relationship between variables and the choice of prior is such that the resulting ESS prior distribution is close to Jeffrey’s prior. The software infers causal relationship according to the statistical dependence under some additional assumptions concerning latent variables.
- The so-called causal structure assumes that dependencies between variables are due to causal relationships between variables in the model.

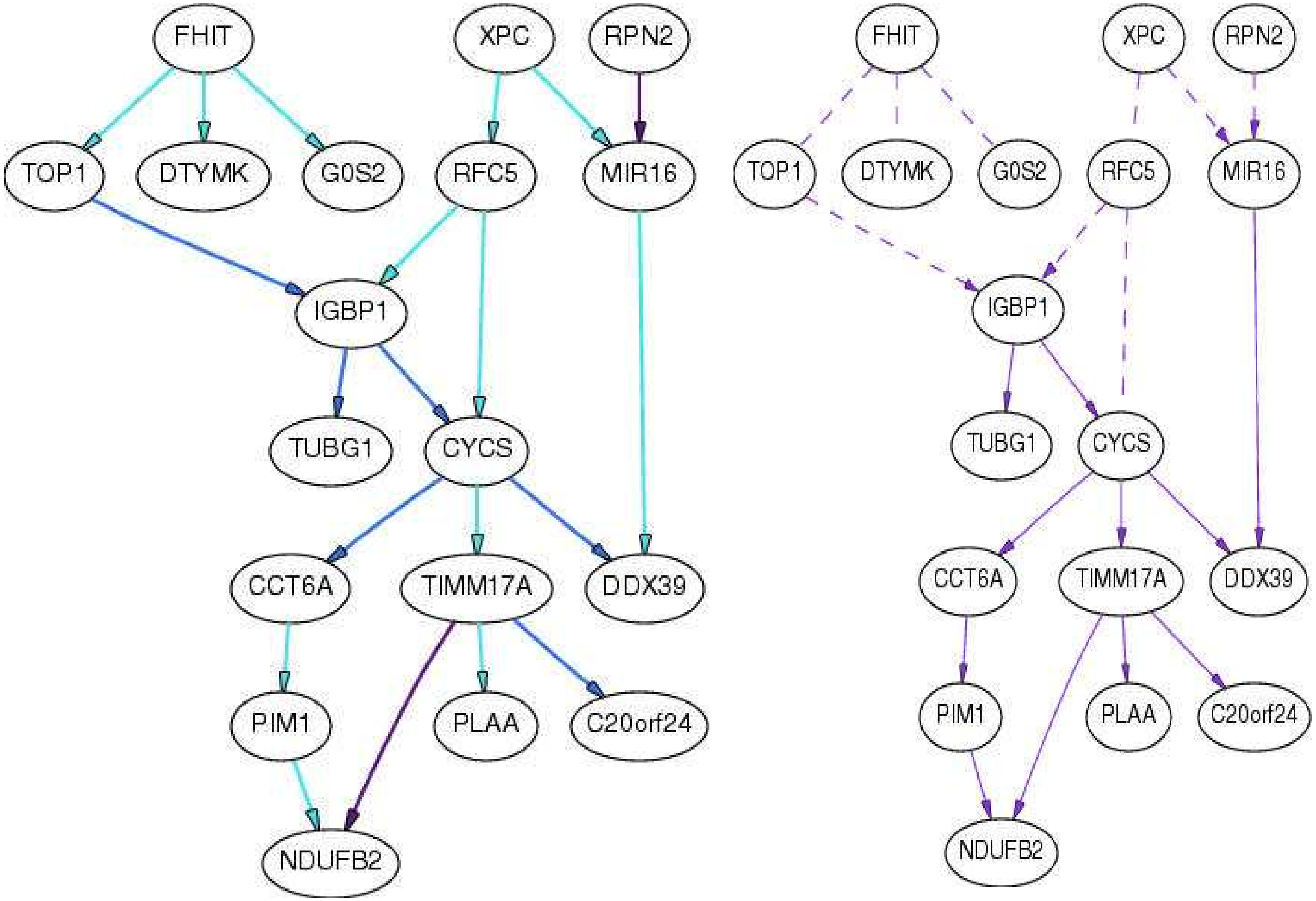
Results

- Cluster analysis shows that the dendrogram (not shown) differ somewhat from the earlier report (Morley et al**.** Nature 2004, **430:** 743-74) possibly due to difference in sample sizes.
- The following genes are independent of any other genes in the model: NFYC, LSM3, RAN, VAMP2, RAP80, INPP5A, STC2, SNRPB. Edges TIMM17A to NDUFB2 and RPN2 to MIR16 are very strong and removing any of them would result in a model with probability less than one millionth that of the original model.

Strength of the dependency

Edge set one			Edge set two		
Vertex 1	Vertex 2	Ratio	Vertex 1	Vertex 2	Ratio
TOP1	IGBP1	436736	CYCS	TIMM17A	102
TIMM17A	C20orf24	201449	RFC5	CYCS	92
CYCS	CCT6A	89880	FHIT	TOP1	61
IGBP1	TUBG1	16221	PIM1	NDUFB2	41
CYCS	DDX39	9248	RFC5	IGBP1	31
IGBP1	CYCS	4388	FHIT	DTYMK	17
			XPC	MIR16	15
			XPC	RFC5	9.96
			FHIT	G0S2	5.85
			MIR16	DDX39	3.58
			TIMM17A	PLAA	3.57
			CCT6A	PIM1	3.31

- Removing any of the edges (Vertex 1 to Vertex 2) in edge set one from the chosen model would decrease the probability of the model to less than one thousandth the probability of the original model, while removing any of the edges in edge set two decreases the probability of the model (exact ratio listed).



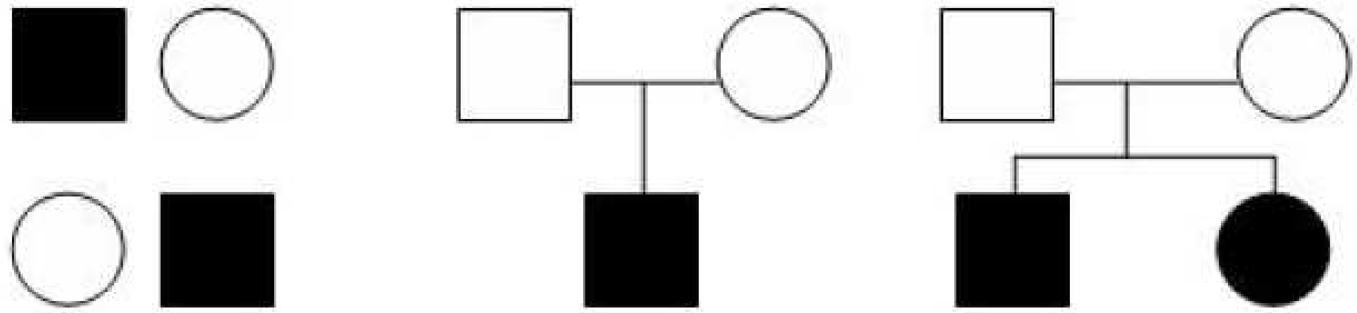
Left. Importance of the dependencies. **Right**. Solid arc has direct causal influence (direct meaning that causal influence is not mediated by any other variable that is included in the study). Dashed arc indicates there are two possibilities, but we do not know which holds. Dashed line without any arrow heads indicates there is a dependency but we do not know the reciprocal dependence.

Summary of Bayesian network analysis

- The series of papers on these data stress the importance of Intermediate phenotypes. Without a priori biological hypothesis, it serves as an exploratory tool for subsequent confirmatory analysis.
- This particular analysis highlights the potential usefulness of pathway analysis but with great work to be done.
- An apparent limitation of this work, though not uncommon in gene-expression studies, is the relatively small sample size used. To fully elucidate the biological pathways involved may be difficult, as for instance CYCS is involved in six pathways according to http://escience.invitrogen.com/ipath/.
- Statistical robustness and biological interpretability remain as the two main challenges for Bayesian network analyses, to which replication, bootstrap and benchmarking have been proposed.
- Our inference of gene networks also exploits the covariance structure of the data, like structural equation modelling, but is exploratory or hypothesis-generating rather than confirmatory or hypothesis-driven. A number of other software systems are of interest, e.g. ASIAN (a web-based regulatory network framework, http://eureka.cbrc.jp), deal.

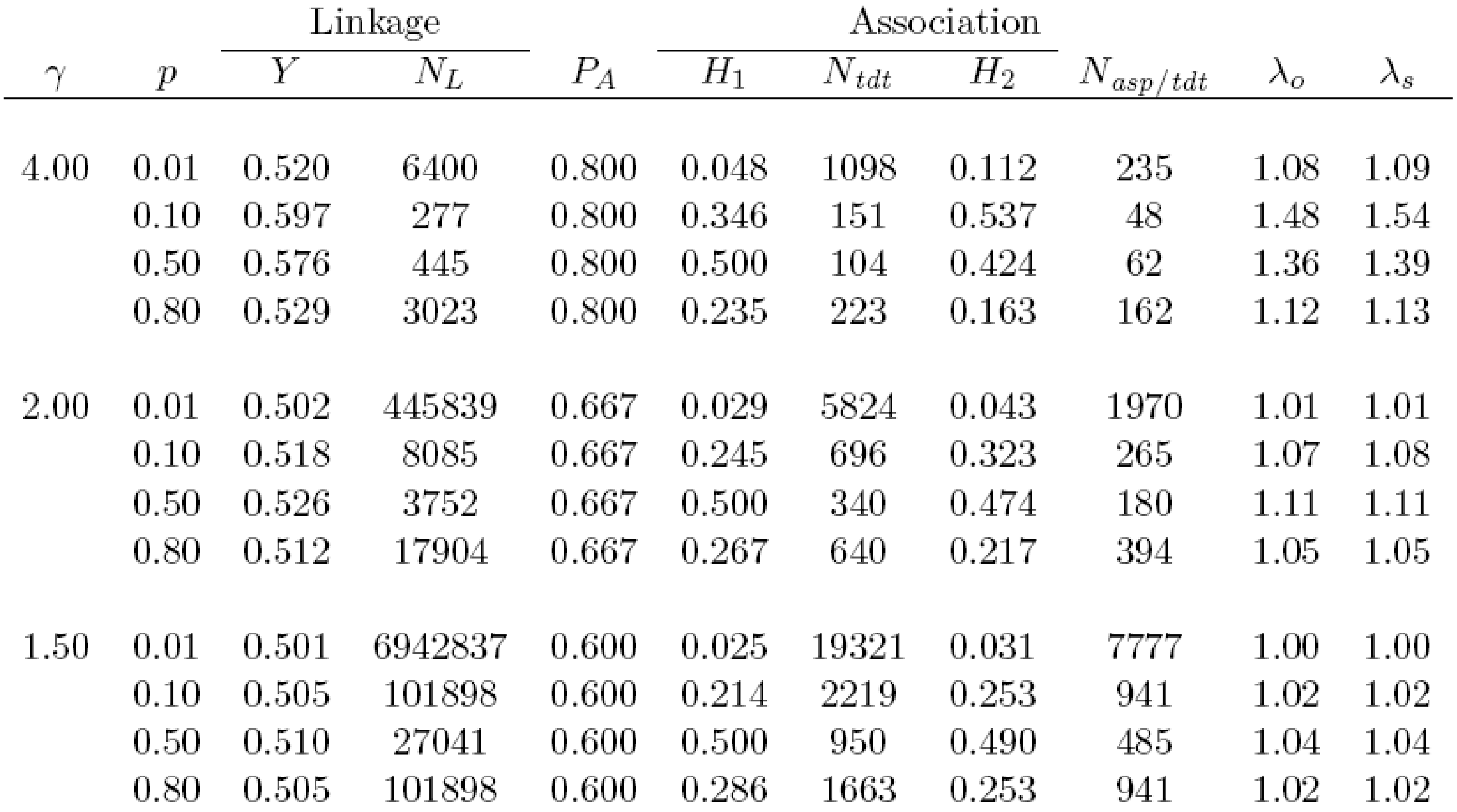
II.2 Genomewide association Studies (GWAS) of Anthropometric and Related Traits

Designs



Three common genetic association designs involving unrelated individuals (left), nuclear families with affected singletons (middle) and affected sib-pairs (right). Males and females are denoted by squares and circles with affected individuals filled with black and unaffected individuals being empty.

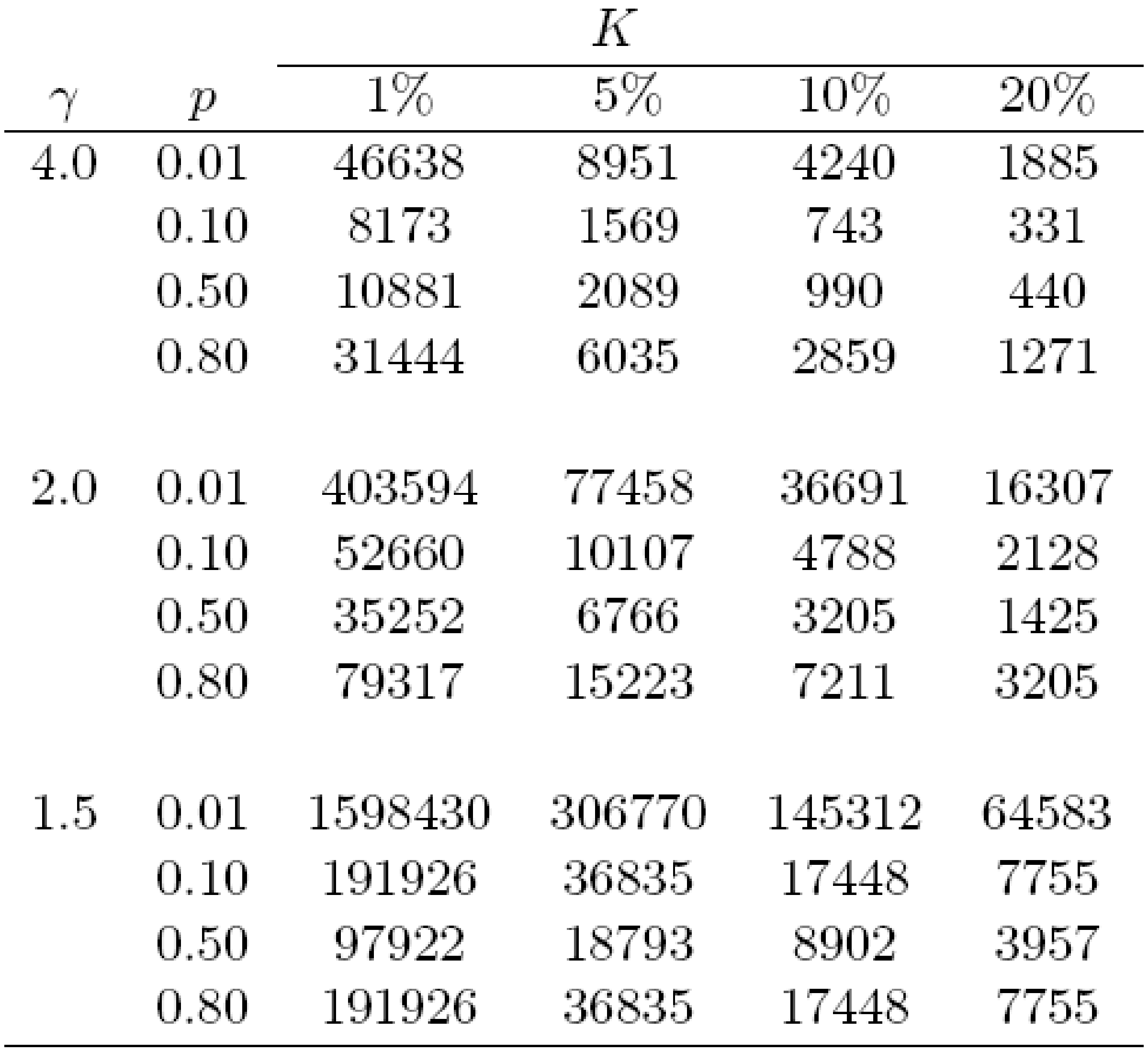
Results for family designs



		Linkage			Association					
γ	p	Y	N_L	P_A	H_1	N_{tdt}	H_2	$N_{asp/tdt}$	λ_o	λ_s
4.00	0.01	0.520	6400	0.800	0.048	1098	0.112	235	1.08	1.09
	0.10	0.597	277	0.800	0.346	151	0.537	48	1.48	1.54
	0.50	0.576	445	0.800	0.500	104	0.424	62	1.36	1.39
	0.80	0.529	3023	0.800	0.235	223	0.163	162	1.12	1.13
2.00	0.01	0.502	445839	0.667	0.029	5824	0.043	1970	1.01	1.01
	0.10	0.518	8085	0.667	0.245	696	0.323	265	1.07	1.08
	0.50	0.526	3752	0.667	0.500	340	0.474	180	1.11	1.11
	0.80	0.512	17904	0.667	0.267	640	0.217	394	1.05	1.05
1.50	0.01	0.501	6942837	0.600	0.025	19321	0.031	7777	1.00	1.00
	0.10	0.505	101898	0.600	0.214	2219	0.253	941	1.02	1.02
	0.50	0.510	27041	0.600	0.500	950	0.490	485	1.04	1.04
	0.80	0.505	101898	0.600	0.286	1663	0.253	941	1.02	1.02

Y=probability of allele sharing; P_A=probability of transmitting disease allele A; H_1, H_2=proportions of heterozygous parents

Case-control design



		K			
γ	p	1%	5%	10%	20%
4.0	0.01	46638	8951	4240	1885
	0.10	8173	1569	743	331
	0.50	10881	2089	990	440
	0.80	31444	6035	2859	1271
2.0	0.01	403594	77458	36691	16307
	0.10	52660	10107	4788	2128
	0.50	35252	6766	3205	1425
	0.80	79317	15223	7211	3205
1.5	0.01	1598430	306770	145312	64583
	0.10	191926	36835	17448	7755
	0.50	97922	18793	8902	3957
	0.80	191926	36835	17448	7755

A variation: case-cohort design

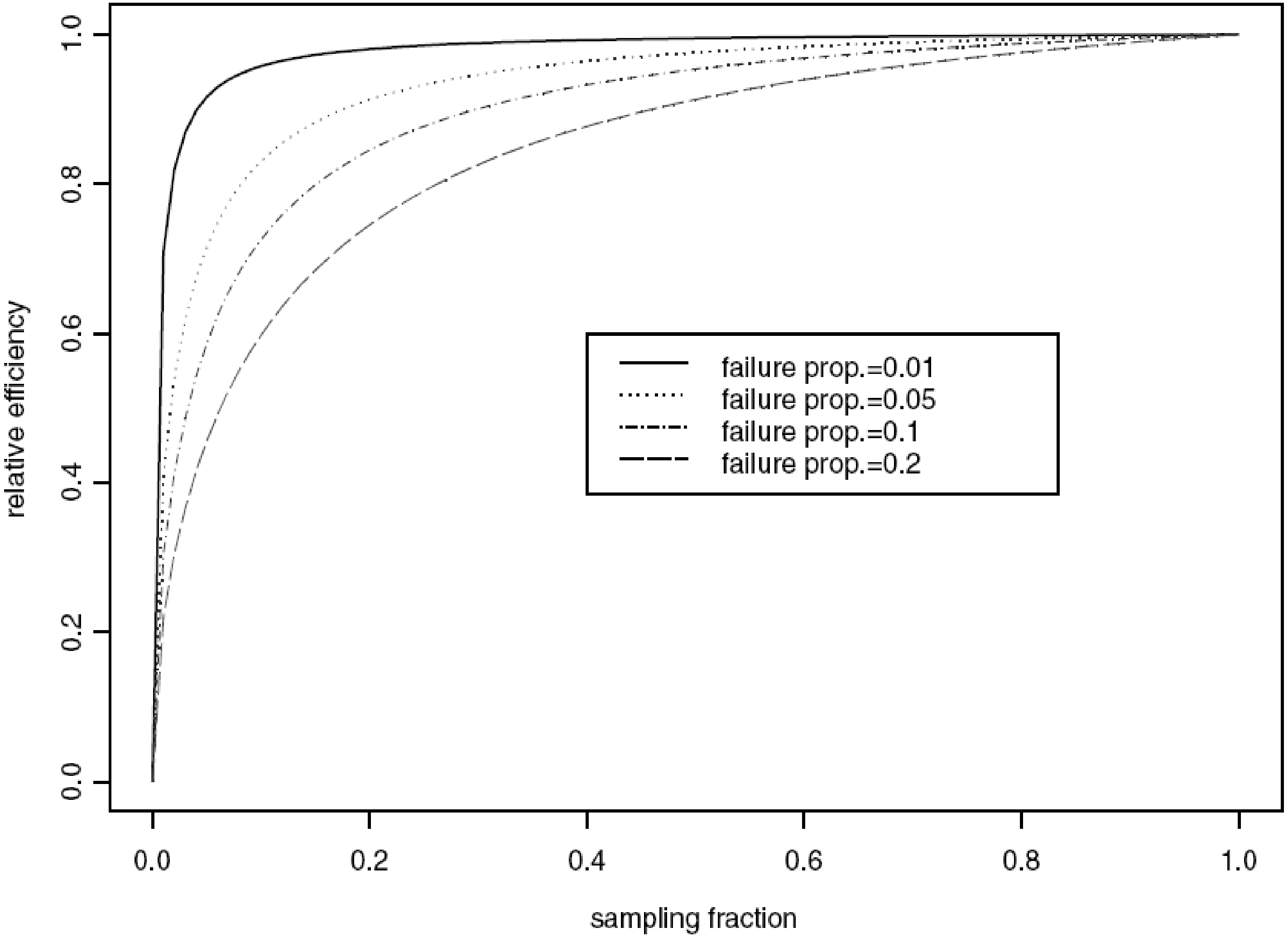
- Assumption: the censoring distributions in two groups, small but moderate number of failures in the full cohort and no ties.
- The power can be obtained from the following expression,

$$\Phi(Z_\alpha + m^{0.5}\theta\sqrt{p_1 p_2 p_D/q + (1-q)p_D})$$

- where α is the significance level, θ is the log-hazard ratio for two groups, p_j, j=1,2 are the proportion of the two groups in the population, m is the total number subjects in the subcohort, p_D is the proportion of the failures in the full cohort, q is the sampling fraction of the subcohort.
- Alternatively, sample size can be obtained (n is the size of cohort),

$$m = nBp_D/(n-B(1-p_D)) \quad B = (Z_{1-\alpha}+Z_\beta)^2/(\theta^2 p_1 p_2 p_D))$$

Relative efficiency of the case-cohort design compared to the full cohort



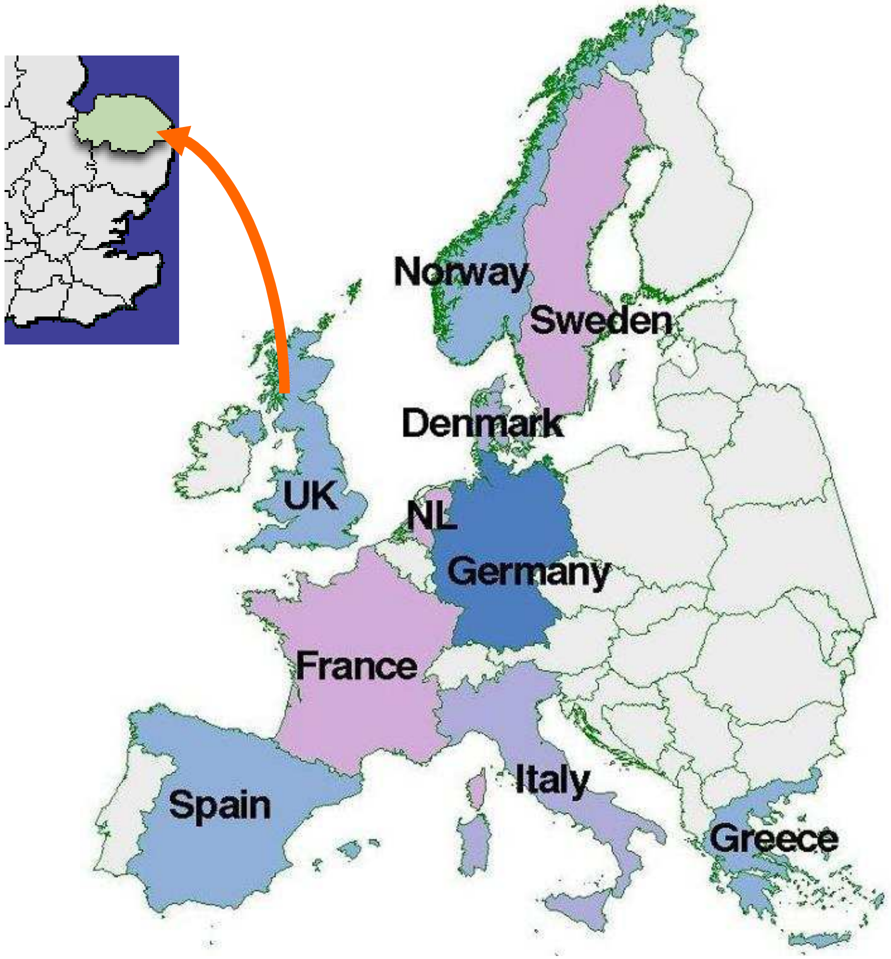
The EPIC Study

The European Prospective Investigation into Cancer and Nutrition (EPIC) is coordinated by Dr Elio Riboli, Head of the Division of Epidemiology, Public Health and Primary Care at the Imperial College London.

EPIC was designed to investigate the relationships between diet, nutritional status, lifestyle and environmental factors and the incidence of cancer and other chronic diseases. EPIC is the largest study of diet and health ever undertaken, having recruited over half a million (520,000) people in ten European countries: Denmark, France, Germany, Greece, Italy, The Netherlands, Norway, Spain, Sweden and the United Kingdom.

The EPIC-Norfolk study

EPIC-Norfolk participants are men and women (based on over 30,000 people) who were aged between 45 and 74 when they joined the study, who lived in Norwich and the surrounding towns and rural areas. They have been contributing information about their diet, lifestyle and health through questionnaires, and through health checks carried out by EPIC nurses.



The case-cohort design for EPIC-Norfolk

- It originally followed case-control design (e.g., WTCCC with seven cases and common controls) with 3425 cases and 3400 controls.
  - It is potentially more powerful.
  - Controls are selected.
- It has then been changed into case-cohort design, in which cases are defined to be individuals whose BMI above 30 and controls are a random sample (subcohort) of the EPIC-Norfolk cohort which includes obese individuals.
  - The subcohort is representative of the whole population and allows for a range of traits to be examined.
  - The analysis is potentially more involved but established.

Power and sample size

- It started with assessment of how the power is compromised relative to the original case-control design.
- This was followed by power/sample size calculation using methods established by Cai and Zeng (2004) as implemented in an R function, noting a number of assumptions.
- More practically, it was also envisaged that a proper representative sample of a total of 25,000 individuals would be 10%; the subcohort is then approximately 2,500.
- The total sample was split between two stages.
- Affymetrix 500K data were available for 3850 individuals
- Illumina 317K -- it came at a later time and the quality of data appears to be poor?
- The focus has therefore been Affy500K, but with a possible comeback.

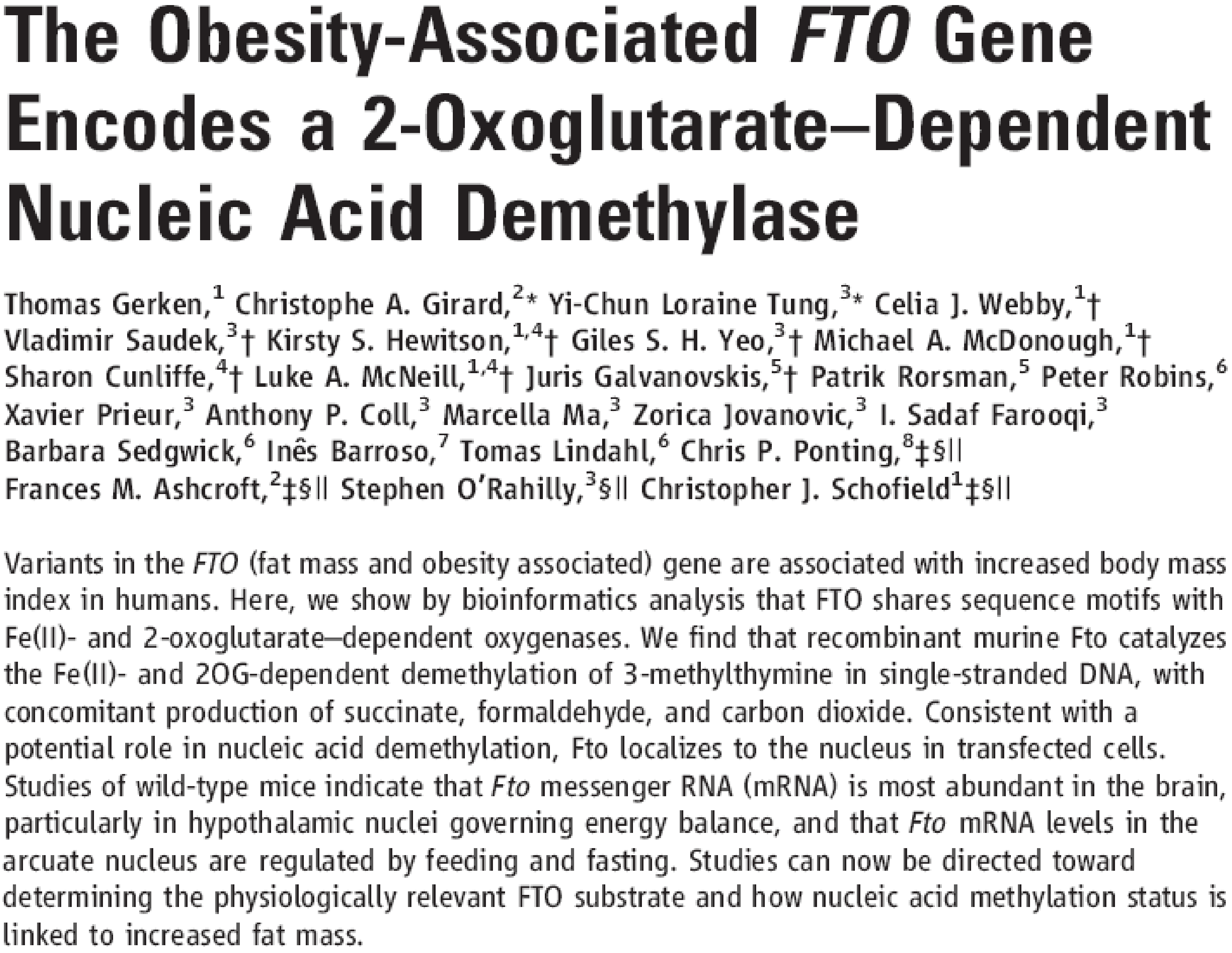
Analysis

- Linux clusters are now ready for comprehensive analyses.
- Linux/awk script is light and appears to be more transparent than Perl, Java which is more professional.
- awk proves very useful and can be transformed to Perl. In fact, any statistical package which processes data elements would be less efficient. An example is the transformation of long, wide, transposed format noted earlier.
- They call C/C++ programs such as IMPUTE/SNPTEST.
- SAS is still useful for data preparation, and in a sense less professional than DBMS such as Oracle but enjoys a large user community and has facility for data analysis.
- SAS 9.2 PROTO procedure is yet to be explored.

Findings

- Frayling et al. Science 316:889-894, 2007
- Sandhu MS et al. Lancet 371:483-91, 2008
- Weedon et al. Nat Genet 40:575-583
- Loos et al. Nat Genet 40:768-775, 2008
- Prokopenko et al. Nat Genet (in press)
- Willer et al. Nat Genet (in press)

- … blood pressure/HT, Alcohol, IHD, BMD, FEV1



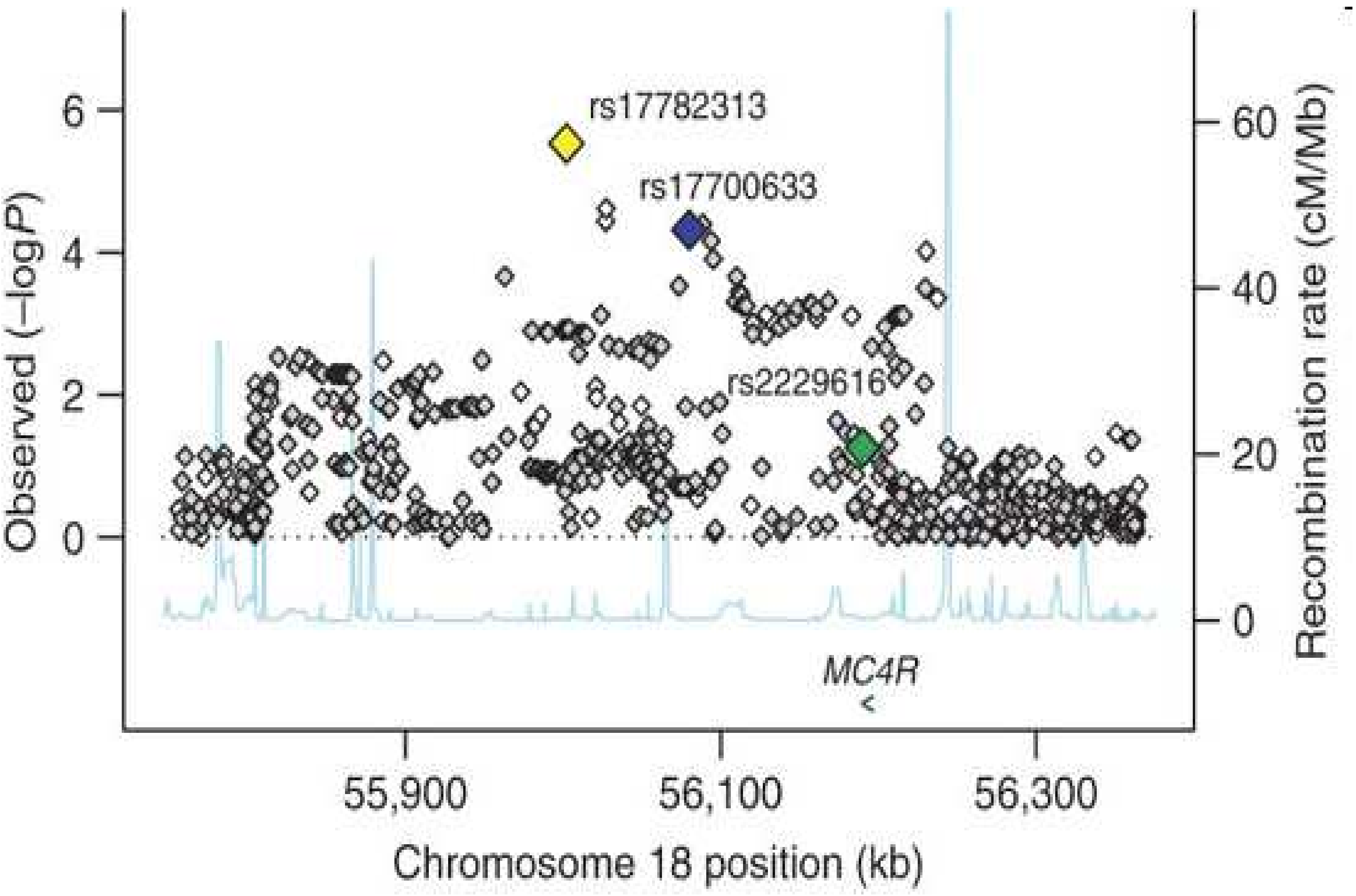
The Obesity-Associated *FTO* Gene Encodes a 2-Oxoglutarate–Dependent Nucleic Acid Demethylase

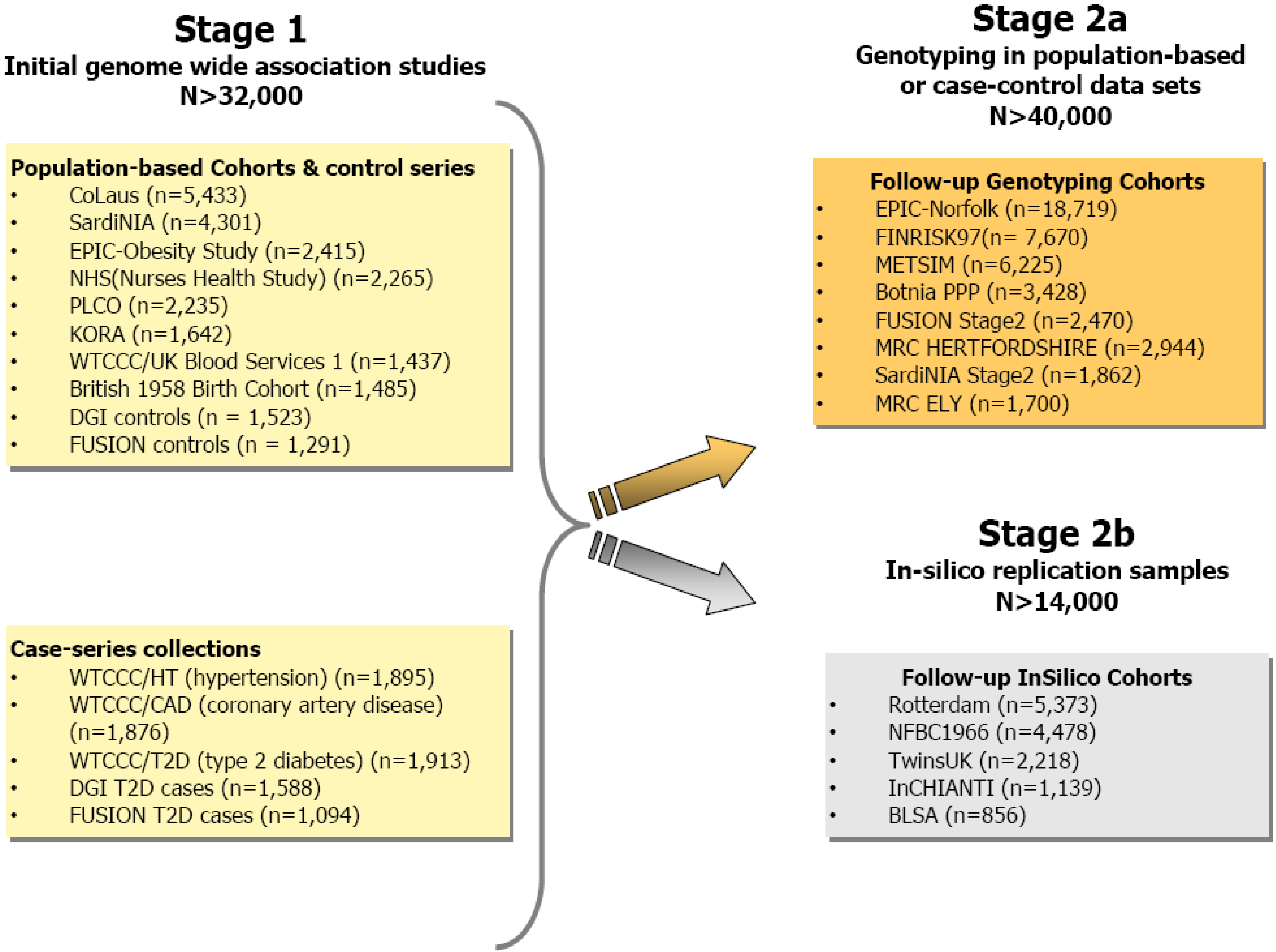
Thomas Gerken,[1] Christophe A. Girard,[2]* Yi-Chun Loraine Tung,[3]* Celia J. Webby,[1]† Vladimir Saudek,[3]† Kirsty S. Hewitson,[1,4]† Giles S. H. Yeo,[3]† Michael A. McDonough,[1]† Sharon Cunliffe,[4]† Luke A. McNeill,[1,4]† Juris Galvanovskis,[5]† Patrik Rorsman,[5] Peter Robins,[6] Xavier Prieur,[3] Anthony P. Coll,[3] Marcella Ma,[3] Zorica Jovanovic,[3] I. Sadaf Farooqi,[3] Barbara Sedgwick,[6] Inês Barroso,[7] Tomas Lindahl,[6] Chris P. Ponting,[8]‡§|| Frances M. Ashcroft,[2]‡§|| Stephen O'Rahilly,[3]§|| Christopher J. Schofield[1]‡§||

Variants in the *FTO* (fat mass and obesity associated) gene are associated with increased body mass index in humans. Here, we show by bioinformatics analysis that FTO shares sequence motifs with Fe(II)- and 2-oxoglutarate–dependent oxygenases. We find that recombinant murine Fto catalyzes the Fe(II)- and 2OG-dependent demethylation of 3-methylthymine in single-stranded DNA, with concomitant production of succinate, formaldehyde, and carbon dioxide. Consistent with a potential role in nucleic acid demethylation, Fto localizes to the nucleus in transfected cells. Studies of wild-type mice indicate that *Fto* messenger RNA (mRNA) is most abundant in the brain, particularly in hypothalamic nuclei governing energy balance, and that *Fto* mRNA levels in the arcuate nucleus are regulated by feeding and fasting. Studies can now be directed toward determining the physiologically relevant FTO substrate and how nucleic acid methylation status is linked to increased fat mass.

Science 318:1469-1472, 2007

Association of *MC4R* and obesity (Loos et al. 2008)





Stage 1

Initial genome wide association studies

N>32,000

Population-based Cohorts & control series

- CoLaus (n=5,433)
- SardiNIA (n=4,301)
- EPIC-Obesity Study (n=2,415)
- NHS(Nurses Health Study) (n=2,265)
- PLCO (n=2,235)
- KORA (n=1,642)
- WTCCC/UK Blood Services 1 (n=1,437)
- British 1958 Birth Cohort (n=1,485)
- DGI controls (n = 1,523)
- FUSION controls (n = 1,291)

Case-series collections

- WTCCC/HT (hypertension) (n=1,895)
- WTCCC/CAD (coronary artery disease) (n=1,876)
- WTCCC/T2D (type 2 diabetes) (n=1,913)
- DGI T2D cases (n=1,588)
- FUSION T2D cases (n=1,094)

Stage 2a

Genotyping in population-based or case-control data sets

N>40,000

Follow-up Genotyping Cohorts

- EPIC-Norfolk (n=18,719)
- FINRISK97(n= 7,670)
- METSIM (n=6,225)
- Botnia PPP (n=3,428)
- FUSION Stage2 (n=2,470)
- MRC HERTFORDSHIRE (n=2,944)
- SardiNIA Stage2 (n=1,862)
- MRC ELY (n=1,700)

Stage 2b

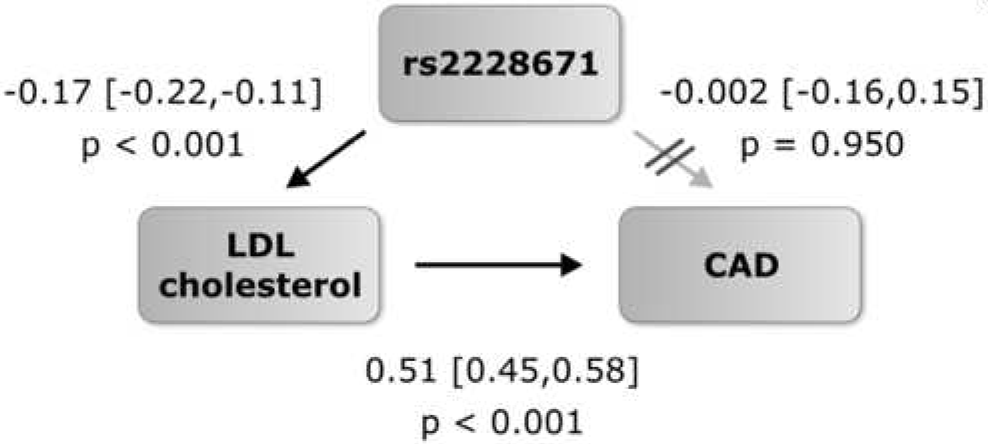
In-silico replication samples

N>14,000

Follow-up InSilico Cohorts

- Rotterdam (n=5,373)
- NFBC1966 (n=4,478)
- TwinsUK (n=2,218)
- InCHIANTI (n=1,139)
- BLSA (n=856)

Mendelian randomization

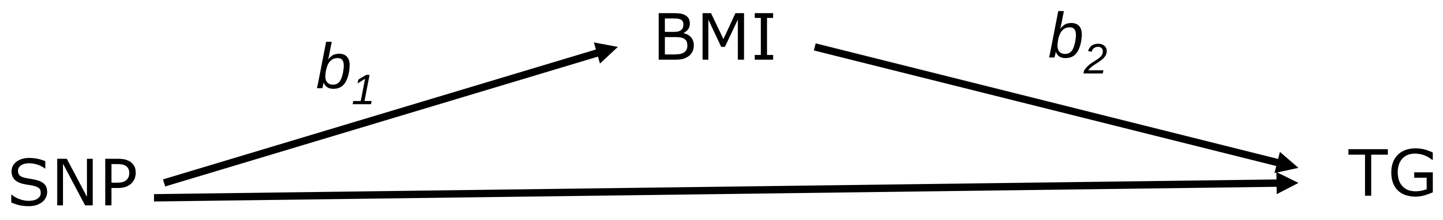


Causal relationship between LDL-C associated with rs2228671 and CAD (Mendelian Randomisation). In the structural equation model, carriage of the T allele at rs2228671 leads to lower LDL-C levels, and higher LDL-C levels lead to an increased risk of CAD. Given this, there is no additional direct path from rs2228671 to CAD risk, indicating that the functional pathway between the genetic variant at the *LDLR* gene locus and risk of CAD is through changes in LDL-C.

Linsel-Nitschke et al. (2008) PLoS ONE, 3(8):e2986

Specific analysis

- Which trait MC4R has effect on?
- Interpretation of mediation
  - Path analysis − shows mainly on BMI and not others
  - Error propagation as appropriate for meta-analysis



As is the case with FTO and T2D, the indirect effect (IE) from MC4R SNP to TG via BMI is b_1b_2, with $SE(IE) \approx b_1SE(b_2)$

III. Gene characterization

III.1 *APOE* and cognitive function (Zhao et al. 2005)

The data (dubbed as the "general knowledge")

Tests	Description	Duration
Memory	An audiotaped list of twenty single or double syllable words at two-second intervals and asked to recall them in writing	2 Min
AH4	A test of 65 verbal and numeric items to be completed	10 Min
Mill Hill	A test of thirty-three multiple choices from groups of six words	10 Min
"S" words	To recall in writing as many words beginning with the letter "S"	1 Min
Animals	To recall in writing as many "animal" words	1 Min

The AH4 tests for fluid ability which is associated with reasoning and induction and Mill Hill tests for crystallised ability which is associated with accumulation of knowledge and vocabulary.

	S3 Age	MEM	AH4	MH	SWORDS	ANIMALS
Men						
Low	1	3.63 (0.31)	4.00 (0.60)	1.23 (0.19)	3.00 (0.36)	2.71 (0.31)
	2	3.25 (0.27)	5.08 (0.58)	1.24 (0.20)	2.16 (0.29)	2.33 (0.26)
	3	2.69 (0.32)	5.57 (0.54)	1.00 (0.28)	2.57 (0.39)	2.49 (0.29)
	4	3.16 (0.30)	3.56 (0.55)	1.24 (0.25)	1.79 (0.30)	1.69 (0.30)
Medium	1	2.10 (0.13)	2.17 (0.33)	0.44 (0.11)	1.07 (0.19)	1.56 (0.18)
	2	1.91 (0.12)	1.67 (0.31)	0.22 (0.11)	1.21 (0.20)	1.59 (0.19)
	3	1.87 (0.16)	0.83 (0.43)	0.26 (0.15)	0.49 (0.27)	1.08 (0.28)
	4	1.05 (0.13)	0.18 (0.40)	0.07 (0.17)	-0.08 (0.20)	0.43 (0.18)
High	1	0.30 (0.17)	-1.28 (0.42)	-0.35 (0.12)	-0.95 (0.27)	0.33 (0.23)
	2	0.27 (0.18)	-1.05 (0.31)	-0.32 (0.12)	-1.01 (0.29)	-0.60 (0.26)
	3	-0.25 (0.23)	-2.14 (0.45)	-0.16 (0.13)	-1.62 (0.36)	-0.28 (0.30)
	4	-1.08 (0.28)	-1.84 (0.47)	-0.32 (0.13)	-3.03 (0.44)	-1.09 (0.33)
Women						
Low	1	4.33 (0.47)	7.33 (1.09)	2.21 (0.66)	3.75 (0.96)	3.88 (1.91)
	2	3.33 (0.43)	4.74 (1.09)	2.30 (0.64)	1.86 (0.63)	3.91 (0.74)
	3	1.93 (0.37)	4.07 (1.05)	1.66 (0.41)	1.32 (0.52)	2.46 (0.48)
	4	2.16 (0.36)	3.77 (0.84)	1.07 (0.37)	1.89 (0.38)	1.78 (0.29)
Medium	1	2.08 (0.32)	2.62 (0.73)	1.08 (0.22)	1.71 (0.40)	2.65 (0.36)
	2	1.90 (0.30)	2.41 (0.62)	0.24 (0.24)	0.96 (0.41)	1.80 (0.30)
	3	1.18 (0.32)	2.23 (0.64)	0.18 (0.25)	0.84 (0.65)	1.32 (0.39)
	4	1.78 (0.26)	-0.16 (0.66)	0.21 (0.22)	0.58 (0.42)	0.52 (0.29)
High	1	0.31 (0.31)	-0.84 (0.51)	-0.05 (0.18)	-0.31 (0.41)	0.43 (0.50)
	2	0.15 (0.29)	-2.11 (0.84)	-0.38 (0.20)	-0.96 (0.45)	-0.02 (0.45)
	3	-0.66 (0.35)	-1.84 (0.87)	-0.47 (0.29)	-1.33 (0.62)	-1.52 (0.94)
	4	-0.51 (0.48)	-2.35 (1.07)	-0.39 (0.36)	-3.06 (1.03)	-1.76 (0.71)

Participants scoring low at phase 3 showed an overall improvement, while those scoring high at phase 3 showed decline, indicating regression towards the mean!

Tests of cohort and practice effect

Ideally cohort effects can be tested using two independent samples at different time points; it may reflect sociohistorical changes on the study samples involved (Jacqmin-Gadda et al. 1997);

The differentiation of practice effect is also knowingly difficult (e.g. Clarke 2004)

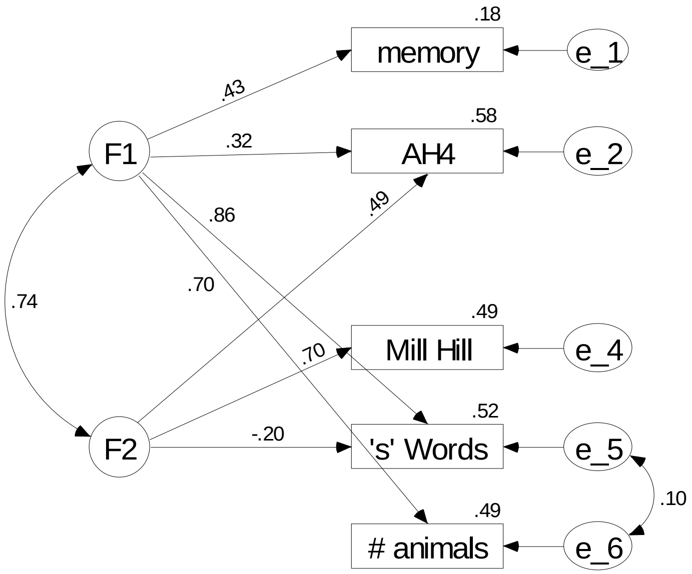
	Phase 5 attenders	Phase 5 non-attenders
Phase 3 attenders	*a*	*b*
Phase 3 non-attenders	*c*	*d*

Test of these two effects are possible since cell *a* accounts for 40% of the total sample at phase 3, while cells *b* and d together account for all the sample at phase 5

- Matched with age between *b* and *c* to test for cohort effect
- Comparison between *c* and *d* for practice effect

In men, MEM, AH4, “S” WORDS all with p<0.0001 and getting worse with β=-0.40, -1.48 and -0.59 but not MH and ANIMALS; the negative signs indicate higher scores in the younger sub-cohort that started at phase 5.

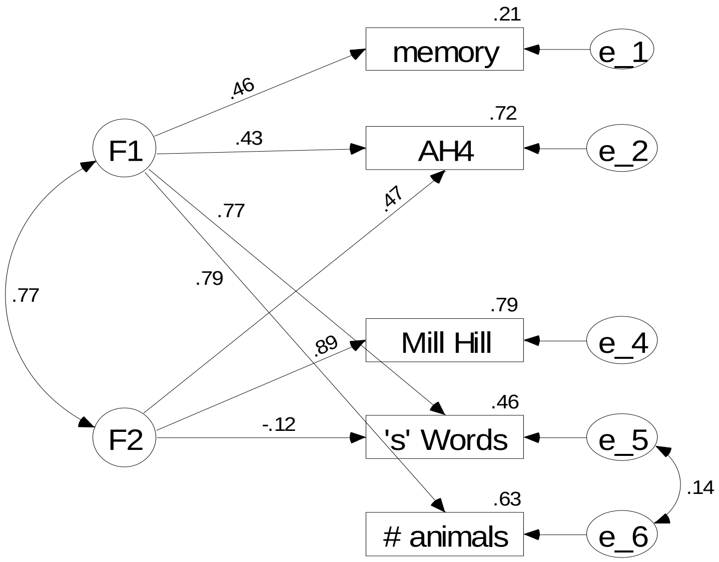
In women, MEM and MH with p=0.0086 and 0.021 with β=-0.35 and 0.56.



Chi-square=3.752 (2 df) p=.153 RMSEA=.012 CFI=1.000

Two-factor model

Men



Chi-square=3.752 (2 df) p=.153 RMSEA=.012 CFI=1.000

Two-factor model

Women

It appears the effect is on the memory; we had a sporadic finding on memory in the data set as well but it was rather sensitive to a small number of individuals. Can we export the score as a new phenotype? Actually there is still a correlation between the latent factors F1 and F2.

III.2 A GAW16-FHS data analysis

- Framingham Heart Study (FHS) under the direction of National Heart, Lung, and Blood Institute (NHLBI) which began in 1948 with the recruitment of adults from the town of Framingham, Massachusetts. At the time, little was known about the general causes of heart disease and stroke, but the death rates for cardiovascular disease (CVD) had been increasing steadily since the beginning of the 20th century and had become an American epidemic. The FHS is now conducted in collaboration with Boston University.
- The data consists of 7130 individuals of the original, the first and the third generation cohorts of which ~7000 individuals have Affymetrix 500K data.
- We can replicate findings on *FTO*, *MC4R*, but not *PCSK1* as reported elsewhere and *INSIG2*, *NYD-SP18* as reported from FHS itself.

IV. Summary

In a nutshell

- It is currently a norm to use HapMap as a resource for data imputation, e.g., from Affymetrix 500K to ~2.5M SNPs
- The upcoming 1000 genome project is expected to play a similar and possibly larger role in the near future, in the sense that it should allow for better characterization of the study sample than that offered by HapMap.
- There is enough to embrace and celebrate for GWAS, for which the study design is crucial.
- Information flow between biology and epidemiology is bidirectional.
- We will need to backtrack techniques that may be seen as somewhat improper.
- However, there is a great deal of certainty that epidemiology will play a significant role in delineating aetiology and variation of human diseases and quantitative traits.

Acknowledgements

- MRC Epidemiology Unit and the GIANT consortium for GWAS
- GAW15 for the gene expression data
- GAW16 organizers and NIH for the Framingham data
- Prof Qihua Tan on aging study of the CEPH gene expression data
- Prof Fuzhong Xue and Shandong University for feasibility of a lecture series
- Dr Wuchun Cao for hosting this event
- You for your attention